C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image

Zhang Chen
Xi'an Jiaotong University
Xi'an, China
1900938761@qq.com

Zhiqiang Tian*
Xi'an Jiaotong University
Xi'an, China
zhiqiangtian@xjtu.edu.cn

Jihua Zhu
Xi'an Jiaotong University
Xi'an, China
zhujh@xjtu.edu.cn

Ce Li
Lanzhou University of Technology
Xi'an, China
xjtulice@gmail.com

Shaoyi Du*
Xi'an Jiaotong University
Xi'an, China
dushaoyi@gmail.com

Abstract

Recently, many excellent weakly supervised semantic segmentation (WSSS) works are proposed based on class activation mapping (CAM). However, there are few works that consider the characteristics of medical images. In this paper, we find that there are mainly two challenges of medical images in WSSS: i) the boundary of object foreground and background is not clear; ii) the co-occurrence phenomenon is very severe in training stage. We thus propose a Causal CAM (C-CAM) method to overcome the above challenges. Our method is motivated by two cause-effect chains including category-causality chain and anatomy-causality chain. The category-causality chain represents the image content (cause) affects the category (effect). The anatomy-causality chain represents the anatomical structure (cause) affects the organ segmentation (effect). Extensive experiments were conducted on three public medical image data sets. Our C-CAM generates the best pseudo masks with the DSC of 77.26%, 80.34% and 78.15% on ProMRI, ACDC and CHAOS compared with other CAM-like methods. The pseudo masks of C-CAM are further used to improve the segmentation performance for organ segmentation tasks. Our C-CAM achieves DSC of 83.83% on ProMRI and DSC of 87.54% on ACDC, which outperforms state-of-the-art WSSS methods. Our code is available at https://github.com/Tian-lab/C-CAM.

1. Introduction

Recently, semantic segmentation [22] is widely studied due to the development of deep learning. Existing paradigm of semantic segmentation is training a model to classify the category of every pixel with abundant pixel-level labeled data. However, the acquirement of pixel-level labels is time-consuming and expensive. Therefore, a new paradigm called weakly supervised semantic segmentation (WSSS) comes out. Different from fully supervised semantic segmentation (FSSS), WSSS utilizes weak annotations, e.g., image-level label, point, scribble and bounding box. Among these weak annotations, image-level label is the easiest way to be obtained. Meanwhile, it is the most challenging one to be used for segmentation. In this paper, we focus on image-level labels for medical image segmentation.

The main problem for WSSS with image-level labels is the lack of location information. Class activation mapping (CAM) methods [7, 14, 24, 30, 33, 44] creatively give convolutional neural network (CNN) locating ability with only image-level labels. However, the CAM could only locate discriminative part of object, which leads awful segmentation performance. Many CAM-based WSSS methods [2, 8, 13, 18, 34, 39] are successively proposed to narrow the gap between WSSS and FSSS. The main idea of these

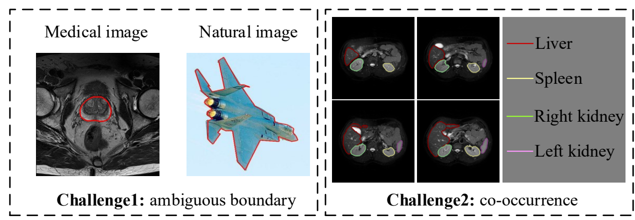

Figure 1. Main challenges of medical images. **Challenge1:** The object boundary for medical image is more ambiguous than natural image. **Challenge2:** Different organs often occur in the same medical image in training stage.

*Corresponding Author

methods is to solve the problem that CAM could not completely cover object. Some methods [2, 8, 13, 18] use CAM to generate seeds and refine the seeds to cover the whole object. Some methods [19, 34, 39, 40] directly generate more accurate saliency maps.

However, most of these CAM-based methods are designed for natural images, which may not work well on medical images. Compared with natural images, medical images have mainly the following two challenges of WSSS based on image-level labels. We intuitively demonstrate the challenges in Fig. 1. i) The boundary of foreground and background is not clear, which makes CAM model hard to classify the category border of foreground and background. ii) The co-occurrence is very severe in medical images in training stage, e.g., different organs always appears together in an abdominal magnetic resonance imaging (MRI) image. However, the co-occurrence is not so severe in natural images. For example, the "people" would not always appear together with "horse", and vice versa. Therefore, CAM model could know which part of an image is "people" with abundant image-level labels. Unfortunately, it is hard for CAM model to activate correct co-occurring organs in one image only according with image-level labels.

Therefore, a causal CAM (C-CAM) method is proposed to overcome the above-mentioned challenges. The C-CAM starts from two causality chains. The first chain is **category causality** $X \rightarrow Y$, which indicates that the image content X (*cause*) affects the classified category Y (*effect*). The second chain is **anatomy causality** $Z \rightarrow S$, which indicates that the anatomical structure Z (*cause*) affects the organ segmentation S (*effect*). In the category-causality chain, we use causal intervention [28] to make C-CAM model focus on the real cause of predicted category. In the anatomy-causality chain, anatomical constraint is integrated to make C-CAM focus on the real cause of object segmentation, which can well solve the co-occurrence problem.

In summary, the main contributions of this paper are three folds:

- We propose C-CAM for WSSS on medical images. The C-CAM generates pseudo segmentation masks with clearer boundaries and more accurate shapes. To the best of our knowledge, C-CAM is the first method to introduce causality into medical image WSSS.

- We integrate two causality chains to cope with the challenges of WSSS for medical images. Category-causality chain is designed to alleviate the problem of ambiguous boundary. Anatomy-causality chain is designed to solve the co-occurrence problem.

- We demonstrate the effectiveness of our method with extensive experiments on three public medical image data sets. Our C-CAM generates pseudo masks with the DSC of 77.26%, 80.34%, and 78.15% respectively on ProMRI, ACDC, and CHAOS data sets. The segmentation performance achieves the DSC of 83.83% $\pm$ 5.14% on ProMRI and 87.54% $\pm$ 7.77% on ACDC, which outperforms state-of-the-art methods.

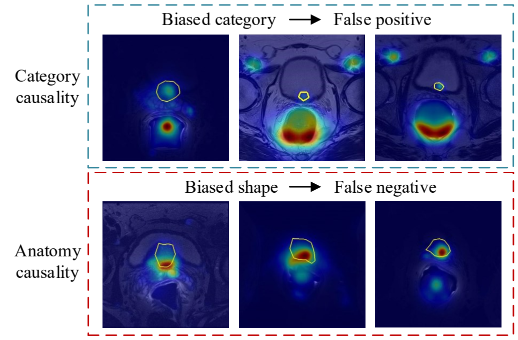

Figure 2. Motivation of causality in medical image WSSS. Category causality (first row): information of biased category cause model activate wrong category of object. Anatomy category (second row): information of biased shape cause model activate inaccurate shape of object. The heatmaps represent saliency maps of CAM. The red color means high value and blue color means low value. The yellow curve represents ground-truth.

2. Related Work

2.1. Weakly Supervised Semantic Segmentation

There are mainly four types of weak labels that are explored in WSSS, including image-level labels [2, 18], points [3], scribbles [20, 37] and bounding boxes [10, 17, 26]. Specially, since the image-level labels are easiest to obtain, most works are designed for image-level WSSS. Our work also focuses on the image-level supervision.

Current image-level supervised WSSS methods are mostly based on CAM technique [44], which could locate discriminative areas with classification model. However, the CAM only activates regions that are highly related to the classified category. Common pipeline of CAM-based method could be divided into three stages. The first stage is to generate seed regions with CAM method. The second stage is to refine seeds regions to generate pseudo masks. The last stage is to train segmentation model with pseudo masks. Many works focus on how to refine seed regions. AffinityNet [2] exploits affinity labels from seed regions and trains an affinity model to refine seed regions. Similarly, BES [8] predicts object boundaries in an explicit way and uses predicted boundaries to revise seed regions. DSRG [13] utilizes the seeded region growing mechanism to gradually refine seed regions. Recently, some researchers design models that directly generate more accurate saliency maps. FickleNet [19] generates more precise saliency maps

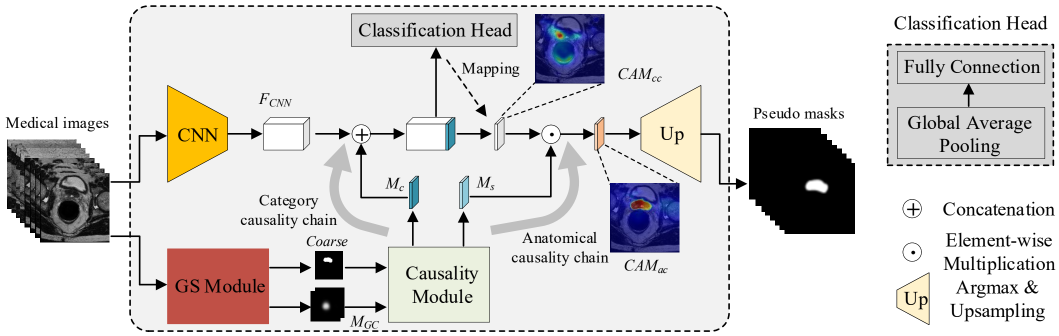

Figure 3. The architecture of our proposed C-CAM. Firstly, a global sampling (GS) module (Sec. 3.2) is designed to generate global context (M_{GC}) and coarse mask. Successively, a causality module (Sec. 3.4) is designed to compute category-causality map M_c and anatomy-causality map M_s. The M_c is then concatenated with CNN features F_{CNN}. The concatenated features $[F_{CNN}, M_c]$ are fed into a classification head in training stage. In inferring stage, CAM_{cc} that represents saliency maps with only category causality is generated by class activation mapping. CAM_{ac} that represents saliency maps with both category causality and anatomy causality is generated by multiplying CAM_{cc} and M_s. Finally, the pseudo masks are generated from CAM_{ac}, which could be used to train a segmentation model in the following full-supervision stage.

by randomly selecting hidden units for a single image. M-CIS [34] exploits cross-image semantic correlations to improve the quality of saliency maps. SEAM [39] uses equivariant regularization to constrain saliency maps of CAM more consistent over rescaling. Wei et al. [40] simply utilizes multi-scale dilated convolution to produce dense and reliable saliency maps. However, these CAM-based methods could not work well on medical images because they do not consider the ambiguous boundary and co-occurrence problems in medical images.

2.2. Anatomical Prior

Incorporating prior knowledge into image segmentation is a useful way to improve performance both for natural image [12, 27, 42] and medical image [11, 25, 31]. In F-SSS scene, current CNN-based methods do not take into account the constraint of output structure since they usually utilize pixel-wise loss functions, e.g. cross-entropy. While a good design of prior can provide better structure constraint [27]. In WSSS scene, prior knowledge is more valuable to make up for the lack of information contained in weak labels [12]. Specially, the priors in medical images have more impact than natural images since objects in medical images naturally have more anatomical information. The anatomical information is inherent like the location of body parts and organs, which is called anatomical prior. Zotti et al. [45] utilize shape prior to aid cardiac MRI segmentation. Mirikharaji et al. [23] design a star shape prior for skin lesion segmentation. Dalca et al. [11] design a generative model for biomedical segmentation, which integrates rich probabilistic anatomical priors. However, existing methods need specialized knowledge or complicated model to utilize anatomical prior. In contrast, our C-CAM extracts anatomical information from the model itself and integrates anatomical prior with an anatomy-causality chain.

2.3. Causality in Computer Vision

Causality has recently been widely used in learning-based computer vision tasks [29, 35, 38, 41, 43]. The introducing of causality to machine learning helps provide better learning and explainable models, since traditional CNN models only take account of association relationship other than causality relationship. Especially, causality plays a more important role in medical imaging. Castro et al. [6] highlight the importance of causality between medical images and their annotations. However, there is no work has been done to apply causality for weakly supervised medical image segmentation as we know. Illuminated by previous excellent works, we introduce causality into weakly supervised semantic segmentation on medical images.

3. Method

3.1. Motivation

We observe that causality plays an important role in medical imaging. The causality for medical image WSSS could be analysed by answering two questions. *Question 1:* why the accuracy of classification model is very high but the activated region of CAM is not accurate? *Question 2:* why the shape of activated region differs far from the ground-truth contour of object? The answer for the first question is that classification model is essentially an association model, which performs well in classification task. However, it does not work in medical image segmentation task. For example, some non-prostate area may has high correlation relationship with prostate in statistical sense. This will lead biased category information that misleads CAM to activate wrong areas that don't have causality relationship with prostate as shown in Fig. 2. The answer for the second question is that current learning-based methods ignore constraint of output structure since they use pixel-wise loss functions. This defect can be remedied with abundant pixel-level labels, while it is obviously not applicable in WSSS scene.

Therefore, two causality chains for WSSS on medical images are proposed to solve the above problems. Category-causality chain is designed to alleviate the problem of ambiguous boundary. Anatomy-causality chain is designed to solve the co-occurrence problem. Fig. 3 shows our C-CAM network structure.

3.2. Global Sampling Module

The saliency map of CAM is not accurate enough for segmentation task. However, it can provide valuable information highly related to category and anatomy for medical image. Therefore, we design a global sampling (GS) module to exploit these valuable information. In this section, the GS module is used to extract global context that contains both category and anatomy information. The GS module is shown as Fig. 4. The training images are directly fed into a pure CAM (P-CAM) model to generate coarse pseudo masks. The P-CAM is a CAM-like model that is composed of a CNN backbone, a classification head, a mapping operation and a upsampling operation. The mapping operation is referred to CAM [44]. In the training stage, only the CNN backbone and classification head are used. In the inferring stage, the mapping operation and a upsampling operation are conducted to generate coarse pseudo masks.

The mapping operation generates saliency maps for each class. This process is defined as a function $f_{p\_cam}(\cdot)$. The GS module finally outputs global context map $M_{GC} \in \mathbb{R}^{C \times H \times W}$, which could be formulated as:

$$M_{GC} = \frac{1}{N} \sum_{k \in N} Up\_Argmax\left(f_{p\_cam}\left(I_k\right)\right), \quad (1)$$

where N denotes the number of training images, $I \in \mathbb{R}^{H \times W \times 3}$ denotes the input images, $Up\_Argmax(\cdot)$ denotes an operation that performs upsampling after argmax, C denotes number of category, H, W denote the height and width of original image size, H', W' denote downsampled size. Specifically, the coarse segmentation mask $Coarse_k = UP\_Argmax\left(f_{p\_cam}\left(I_k\right)\right)$ of every input image is also preserved.

3.3. Causality in medical image WSSS

The key task for WSSS is to generate pseudo mask with accurate category and shape. Our C-CAM starts from two causality chains as shown in Fig. 5. The first chain is **category causality** $X \rightarrow Y$. It indicates that the image content X (*cause*) affects the classified category Y (*effect*) with the disturbing of context confounder C. The second

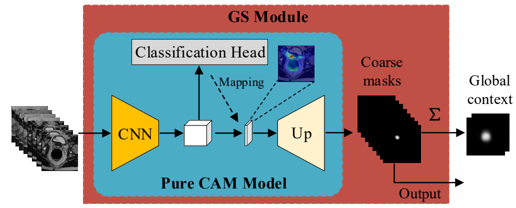

Figure 4. Global Sampling (GS) module.The GS samples all the training data and feeds them into a pure CAM model (P-CAM). The P-CAM generates coarse masks for every training images. In addition, GS module outputs a global context with summarize operation on all coarse masks.

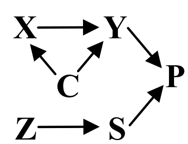

Figure 5. The causal graph of medical image WSSS. X denotes medical image, Y denotes classified category, C denotes context confounder. Z denotes anatomical structure, S denotes shape of segmentation and P denotes pseudo mask.

chain is **anatomy causality** $Z \rightarrow S$, which indicates that the anatomical structure Z (*cause*) affects the shape of segmentation S (*effect*). Therefore, pseudo mask is determined both by category Y and shape S.

3.4. Causality Module

In this section, a causality module is designed as shown in Fig. 6 to improve the accuracy of our P-CAM in a causal manner. As mentioned in Sec. 3.1, the causality module is designed based on two causality chains: category-causality chain and anatomy-causality chain.

Category-Causality Chain. In the category-causality chain, the coarse segmentation mask $Coarse \in \mathbb{R}^{1 \times H \times W}$ and global context map $M_{GC} \in \mathbb{R}^{C \times H \times W}$ are fed into a reshape layer. Two convolution layers are used to project

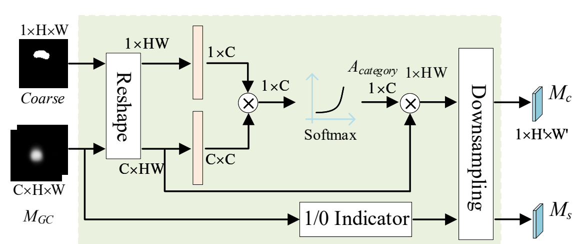

Figure 6. The network structure of causality module. The causality module takes coarse segmentation mask and global context map M_{GC} as input. Finally, the category causality map M_c and the anatomy-causality map M_s are generated respectively for two causality chains.The value of coarse mask is in [0, 1, . . . , C-1], C is the number of categories.

$Coarse$ and M_{GC} into the same space, respectively. A category-aware attention vector $A_{category} \in \mathbb{R}^{1\times C}$ is then computed with the following formulation:

$$A_{category} = softmax\left(\Phi\left(Coarse\right) \times \Theta\left(M_{GC}\right)^{\mathrm{T}}\right), \tag{2}$$

where Φ and Θ represent two convolution operations. Finally, the image-specific category-causality map $M_c \in \mathbb{R}^{1\times H'\times W'}$ is computed as follows:

$$M_c = Down\left(A_{category} \times M_{GC}\right), \tag{3}$$

where $Down\left(\cdot\right)$ is a downsamling operation to make the output M_c could be concatenated with CNN features.

Anatomy-Causality Chain. The shape and boundaries of the targets can be well-capture while the semantic meaning cannot be fully determined, which is later addressed by the anatomical structure information. Especially, for some multi-organ scenes like abdominal scans, CAM_{cc} could even not discriminate left kidney and right kidney since they always co-occur in an image. To this end, an anatomy-causality chain is designed to solve this problem.

In the anatomy-causality chain, a 1/0 indicator is designed to represent anatomical information of medical images. Finally, the anatomy-causality map M_s is computed as the following formulation to obtain the possible position of each category:

$$M_S = \begin{cases} 1, & if\ M_{GC} > 0 \\ 0, & else \end{cases}. \tag{4}$$

The M_s is downsampled and multiplied with CAM_{cc} to get final saliency maps CAM_{ac}. Finally, the pseudo segmentation mask S_{pseudo} is formulated as:

$$S_{pseudo} = UP\_Argmax\left(M_S \cdot CAM_{cc}\right). \tag{5}$$

The generated pseudo segmentation mask is used to train a U-Net [32] model in the following full-supervision stage.

4. Experiments

4.1. Dataset

Three medical image data sets for human organ segmentation were used in our experiments. We used only image-level weak supervision for every data set.

ProMRI. This data set is used for prostate segmentation, which contains 172 volumes of T2-weighted transverse MRI. ProMRI is a mixed data set composed of three subsets that are from PROMISE12 [21], ISBI2013 [5] and in-house data [36]. 30 volumes in PROMISE12 test set are used for testing. The remaining 142 volumes are used for training.

ACDC. This is a data set for left ventricular endocardium segmentation task. The ACDC includes 100 cases of cine MRI, which is publicly available on the 2017 Automatic Cardiac Diagnosis Challenge (ACDC) [4]. The 100 cases are randomly divided into two parts. The first part including 75 cases is used for training. The second part including 25 cases is used for testing.

CHAOS. This is a public data set from the challenge of Combined Healthy Abdominal Organ Segmentation (CHAOS) [15]. This data set contains 4 abdominal organs, which are liver, left kidney, right kidney and spleen. The modality of T2 Spectral Pre-Saturation Inversion Recovery (SPIR) is choosed to evaluate our method. The segmentation task provide 20 cases with labeled masks for training and 20 cases without labeled masks for testing.

4.2. Implementation Details

Our work was mainly implemented in Python and the PyTorch framework. All the codes were ran on Ubuntu 16.04.1 platform with 2 NVIDIA GTX 1080Ti GPUs. In the pseudo-masks generation stage, our P-CAM model was firstly trained with all training images including negative samples. The negative samples represent images that contain no organs. Only positive samples were used to train our C-CAM model. Both the two models were optimized by stochastic gradient (SGD) schedule with different initial learning rate, 0.1 for our P-CAM model and 0.001 for C-CAM model. The U-Net model was adopted to train a segmentation model with pseudo segmentation masks. The segmentation model was optimized by the Adam optimizer with a initial learning rate of $5e^{-4}$. We trained our model for 100 epochs for every data set.

4.3. Ablation Studies for C-CAM

P-CAM	AC	CC	Aff	DSC (%)		
				ProMRI	ACDC	CHAOS
✓				69.45	72.01	52.17
✓			✓	73.80	76.10	64.50
✓	✓			71.88	73.80	70.47
✓		✓		75.54	75.67	56.18
✓	✓	✓		76.10	77.26	75.88
✓	✓	✓	✓	**77.26**	**80.34**	**78.15**

Table 1. The ablation study for each part of C-CAM. **AC**: anatomy causality. **CC**: category causality. **Aff**: affinity refine. **DSC**: Dice Similarity Coefficient.

Tab. 1 gives an ablation study of each module in our approach. It shows that both the category causality and anatomy causality improve the accuracy of pseudo masks

	ProMRI	ACDC	CHAOS				
			Liver	Right kidney	Left kidney	Spleen	Avg.
CAM [44]	69.45	72.01	58.59	42.43	47.26	62.34	52.66
GradCAM [33]	46.09	69.81	56.51	31.06	33.29	50.66	42.88
GradCAM++ [7]	64.68	70.21	59.47	35.59	41.21	54.43	47.68
AblationCAM [30]	48.19	64.14	56.38	19.83	40.29	51.66	42.04
EigenCAM [24]	63.91	45.85	58.78	7.91	41.08	53.83	40.40
LayerCAM [14]	63.67	69.84	59.24	35.57	41.33	54.26	47.60
C-CAM(Ours)	**77.26**	**80.34**	**72.68**	**84.75**	**81.00**	**74.16**	**78.15**

Table 2. Evaluation of different CAM-like localization methods on three data sets with Dice Similarity Coefficient (DSC: %).

on three data sets compared with P-CAM (without any improved design). The anatomy causality achieves 2.43% improvement on ProMRI and 1.79% improvement on ACDC. Especially for multi-label segmentation task, like CHAOS, the anatomy causality brings significant performance elevation by 18.3%. The reason is that the co-occurrence phenomenon is very serious for CHAOS as shown in Fig. 7. Traditional CAM model could not effectively activate correct organ areas in an image without anatomical information. In contrast, our C-CAM could accurately distinguish four different organs appeared in the same image. With the integrated of category causality, the generated pseudo masks further achieve 4.22%, 3.46% and 5.41% DSC Similarity Coefficient (DSC [36]) improvement respectively for ProMRI, ACDC and CHAOS data sets. An affinity model is further trained to improve the accuracy of final pseudo segmentation mask as used in [2]. Finally, the generated pseudo segmentation masks achieves the DSC of 77.26%, 80.34% and 78.15% respectively on three data sets.

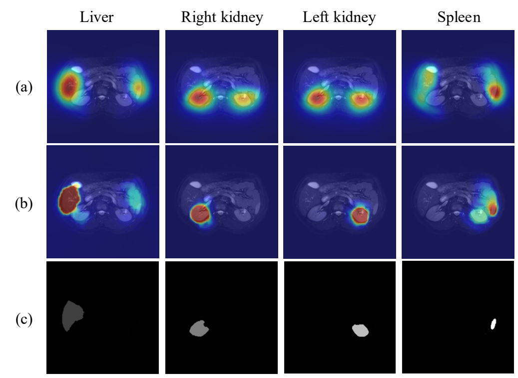

Figure 7. Illustration of co-occurrence that shows the saliency maps of our P-CAM (the first row), our C-CAM (the second row) and the ground-truth (the last row) of four categories. All the presented results correspond to one T2-SPIR image in CHAOS.

4.4. Comparison with other CAM-like methods

Our C-CAM was compared with some CAM-like localization methods, including Grad-CAM [33], Grad-CAM++ [7], Ablation-CAM [30], Eigen-CAM [24] and Layer-CAM [14]. In the experiment, these different CAM-like methods were evaluated with the same trained baseline model used in our C-CAM. All background threshold were tested. All the best DSC results of pseudo masks from different methods were presented, instead of comparing the same threshold for different methods. The evaluation results are shown in Tab. 2. From these results, we find that our C-CAM achieves the best performance of pseudo masks on all three medical image data sets. Especially, our C-CAM performs well on all classes of CHAOS.

4.5. Parameter Sensitivity

The choice of an appropriate background threshold is a basic but critical step to generate pseudo segmentation masks from saliency maps. Extensive experiments were conducted to evaluate the influence of background threshold. Several different CAM-like methods were compared. The accuracy of saliency maps were evaluated with various background thresholds in the range of [0.05, 0.95]. As shown in Fig. 8, most current CAM-like methods are sensitive to different background thresholds. Firstly, the DSC of saliency maps varies a lot with different thresholds for the same CAM-like method. Secondly, the best threshold for DSC is different among these CAM-like methods. Thirdly, the accuracy changes heavily for one threshold value on different data sets. In contrast, our C-CAM is less sensitive to the background thresholds. The DSC of saliency maps from C-CAM could stabilize at high values with background thresholds range from 0.3 to 0.9 as shown in Fig. 8. On the one hand, this would make it easier for us to choose a background threshold. On the other hand, it also indicates the algorithm robustness to background threshold.

4.6. Visualization of saliency maps in C-CAM

Fig. 9 and Fig. 10 give an intuitive illustration of benefits brought from our C-CAM. With the integration of category causality, our C-CAM could well solve the am-

	Methods	Publication&Year	DSC(%) ↑	ASD(mm)↓	MAD(mm)↓
Whole	BES [8]	ECCV(2020)	73.99 ± 6.78	5.16 ± 2.09	5.18± 1.59
	AffinityNet [2]	CVPR(2018)	77.77 ± 6.19	4.04 ± 1.02	4.32 ± 1.33
	SizeLoss [16]	MIA(2019)	81.94 ± 5.66	3.82 ± 1.29	5.00 ± 2.09
	CONTA [43]	NeurlPS(2020)	78.68 ± 5.17	4.16 ± 1.61	4.48 ± 2.57
	IRNet [1]	CVPR(2019)	75.80 ± 5.49	4.72 ± 0.98	5.08 ± 1.24
	ISSOC [9]	PMB(2021)	83.39 ± 5.41	3.80 ± 0.88	3.68 ± 1.21
	P-CAM(Ours)	–	79.02 ± 6.30	3.82± 1.51	3.91 ± 2.01
	C-CAM(Ours)	–	**83.83** ± **5.14**	**3.71** ± **0.78**	**3.36** ± **1.11**
Apex	BES [8]	ECCV(2020)	69.70 ± 10.83	6.14 ± 2.46	6.17 ± 2.73
	AffinityNet [2]	CVPR(2018)	69.22 ± 10.32	6.63 ± 2.84	6.96 ± 4.03
	SizeLoss [16]	MIA(2019)	**73.98** ± **6.39**	3.76 ± 1.78	4.47 ± 2.64
	CONTA [43]	NeurlPS(2020)	72.47 ± 14.86	5.25 ± 1.93	5.25 ± 2.48
	IRNet [1]	CVPR(2019)	63.73 ± 14.11	8.21 ± 3.13	8.39 ± 3.23
	ISSOC [9]	PMB(2021)	68.40 ± 10.48	6.300 ± 2.49	6.090 ± 3.23
	P-CAM(Ours)	–	67.61 ± 13.25	4.29 ± 2.49	4.92 ± 1.87
	C-CAM(Ours)	–	73.00 ± 10.31	**2.03** ± **1.54**	**4.47** ± **1.56**
Base	BES [8]	ECCV(2020)	69.10 ± 11.96	6.70 ± 3.73	7.16 ± 4.94
	AffinityNet [2]	CVPR(2018)	77.81 ± 6.68	4.77 ± 1.90	4.81 ± 2.03
	SizeLoss [16]	MIA(2019)	76.96 ± 9.14	4.75 ± 2.13	5.29 ± 3.72
	CONTA [43]	NeurlPS(2020)	73.89 ± 9.95	5.78 ± 3.37	5.81 ± 4.99
	IRNet [1]	CVPR(2019)	73.73 ± 9.82	6.05 ± 2.62	6.21 ± 3.07
	ISSOC [9]	PMB(2021)	76.82 ± 7.95	4.62 ± 1.52	4.01 ± 1.77
	P-CAM(Ours)	–	73.51 ± 12.24	4.99 ± 2.73	5.79 ± 3.52
	C-CAM(Ours)	–	**85.31** ± **4.76**	**3.22** ± **1.19**	**3.60** ± **1.58**
Mid	BES [8]	ECCV(2020)	79.62 ± 7.33	6.49 ± 3.78	6.96 ± 3.94
	AffinityNet [2]	CVPR(2018)	85.54 ± 5.40	4.13 ± 1.71	4.19 ± 1.86
	SizeLoss [16]	MIA(2019)	86.21 ± 4.43	3.94 ± 1.35	3.92 ± 1.77
	CONTA [43]	NeurlPS(2020)	85.17 ± 5.06	3.93 ± 2.60	3.91 ± 3.07
	IRNet [1]	CVPR(2019)	84.05 ± 4.16	4.94 ± 1.47	5.23 ± 1.64
	ISSOC [9]	PMB(2021)	86.01 ± 5.03	3.93 ± 2.14	3.93 ± 2.17
	P-CAM(Ours)	–	85.09 ± 6.25	4.29 ± 3.03	4.49 ± 3.32
	C-CAM(Ours)	–	**86.40** ± **3.82**	**3.86** ± **1.20**	**3.85** ± **1.33**

Table 3. Comparison of the proposed method with state-of-the-art WSSS methods on ProMRI. The whole-gland and three subregions of prostate volume are compared. The subregions are divided according the prostate size, including apex, mid-gland and base subregions. The prostate size of three subregions: apex $<$ base $<$ mid.

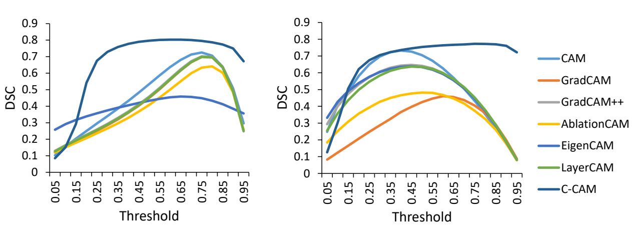

Figure 8. The illustration of sensitivity to the background threshold for different methods. The line charts show accuracy of saliency maps of different methods with various background thresholds on different data sets (left: ProMRI, right: ACDC).

biguous boundary problem. The saliency maps of C-CAM have a clear boundary between foreground and background both on ProMRI and ACDC data sets. In addition, the co-occurrence problem is significant alleviated with the help of anatomy causality as shown in Fig. 9 and Fig. 10. More institutive visualization is shown in Fig. 7. Finally, the saliency maps of C-CAM have fewer error activated areas correspond to unrelated background region, which further verifies the superiority of C-CAM.

4.7. Comparison with other WSSS methods

To further evaluate the effectiveness of our proposed C-CAM, the pseudo segmentation masks were used to train

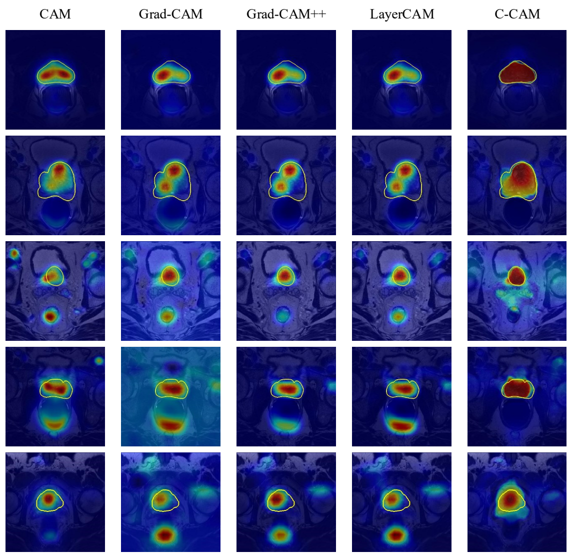

Figure 9. The visualization of saliency maps from different methods for ProMRI. The yellow curve represents ground-truth.

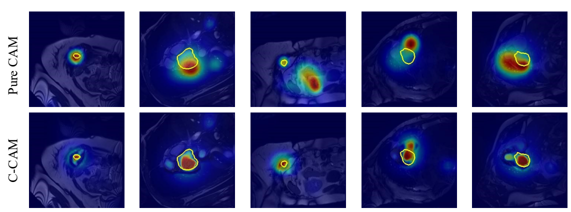

Figure 10. The visualization of saliency maps with/without category causality and anatomy causality for ACDC. The yellow curve represents ground-truth.

	DSC(%) ↑	ASD(mm)↓	MAD(mm)↓
BES [8]	77.53 ± 11.20	2.49 ± 1.41	2.92 ± 2.54
AffinityNet [2]	80.17 ± 8.05	2.28 ± 1.08	2.68 ± 1.97
SizeLoss [16]	80.95 ± 8.57	2.53 ± 1.58	3.31 ± 3.02
IRNet [1]	74.67 ± 14.91	2.79 ± 1.39	3.02 ± 1.86
CONTA [43]	83.51 ± 8.32	1.98 ± 1.68	1.80 ± 0.54
ISSOC [9]	81.65 ± 9.57	2.60 ± 1.66	3.46 ± 3.02
P-CAM(Ours)	75.88 ± 8.70	2.78 ± 1.32	2.89 ± 2.30
C-CAM(Ours)	**87.54 ± 7.77**	**1.62 ± 0.41**	**1.17 ± 0.24**

Table 4. Comparison of the proposed method with state-of-the-art WSSS methods on ACDC.

a U-Net model in full supervision. The final segmentation results of testing data were compared with some other state-of-the-art WSSS methods. Since some other methods are designed for natural images, the codes of these methods were ran on the same data sets used in our experiments for fair comparison. Tab. 3 shows quantitative comparison results. For the whole-gland of prostate, our C-CAM gets the highest DSC of 83.83% with the lowest standard deviation of 5.14%. In terms of the two other metrics average surface distance (ASD) and mean absolute distance (MAD) [36], the C-CAM also achieves the best performance. To verify the performance on different object sizes, the prostate volumes were explicitly compared on three subregions of the whole-gland. The subregions are respectively denoted as apex, base and mid with incremental object size. For three subregions, we can see that our method performs rather well in the mid-gland and base subregions. For the apex subregion, our method achieves slightly lower performance than SizeLoss [16]. The reason is that SizeLoss uses ground-truth to generate weak labels, which avoids error-locating especially for object of small size. However, our C-CAM generate pseudo masks from saliency maps without ground-truth, which may produce large locating error for small objects. The segmentation performance on ACDC data set was also evaluated. Our C-CAM achieves the best performance in terms of all three metrics as shown in Tab. 4. The above experimental results show that our C-CAM significantly improves the performance of segmentation.

5. Conclusion and future work

In this paper, we propose a causality CAM method for WSSS on medical images. Based on the analysis of the causality in medical image WSSS, we design C-CAM that integrates two causal chains to generate accurate pseudo segmentation masks. Category-causality chain is designed to alleviate the problem of ambiguous boundary. Anatomy-causality chain is designed to solve the co-occurrence problem. The generated saliency maps of C-CAM not only have clear boundary between foreground and background, but also keep consistent with anatomical knowledge. The saliency maps of C-CAM outperforms six state-of-the-art CAM-like methods on ProMRI, ACDC and CHAOS data sets. The segmentation network U-Net trained with our pseudo masks achieves state-of-the-art performance on ProMRI and ACDC data sets, which further proves the superiority of our C-CAM. Nevertheless, the proposed C-CAM is hard to segment object with complicated shape. In future work, it is possible to combine a small number of strong annotations and a big number of weak annotations to provide more accurate category and anatomical information.

Acknowledegement

This work was supported in part by NSFC under grant Nos. 62173269 and 61876148, Natural Science Basic Research Plan in Shaanxi Province of China under Grant No. 2022JM-324, Key Research and Development Program of Shaanxi Province of China under Grant No. 2020GXLH-Y-008, Social Science Foundation of Shaanxi Province of China under Grant No. 2021K014.

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 7, 8

[2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 1, 2, 6, 7, 8

[3] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. Whats the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[4] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 5

[5] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370:6, 2015. 5

[6] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020. 3

[7] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 1, 6

[8] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 1, 2, 7, 8

[9] Zhang Chen, Zhiqiang Tian, Yaoyue Zheng, Xiangyu Si, Xulei Qin, Zhong Shi, and Shuai Zheng. Image-level supervised segmentation for human organs with confidence cues. *Physics in Medicine & Biology*, 66(6):065018, 2021. 7, 8

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015. 2

[11] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 3

[12] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. *Advances in Neural Information Processing Systems*, 32:6586–6597, 2019. 3

[13] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018. 1, 2

[14] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Trans. Image Process.*, 30:5875–5888, 2021. 1, 6

[15] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[16] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54:88–99, 2019. 7, 8

[17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2

[18] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 1, 2

[19] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 2

[20] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 2

[21] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 5

[22] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[23] Zahra Mirikharaji and Ghassan Hamarneh. Star shape prior in fully convolutional networks for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 737–745. Springer, 2018. 3

[24] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 1, 6

[25] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P ORegan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017. 3

[26] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015. 2

[27] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015. 3

[28] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2

[29] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869, 2020. 3

[30] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020. 1, 6

[31] Hariharan Ravishankar, Rahul Venkataramani, Sheshadri Thiruvenkadam, Prasad Sudhakar, and Vivek Vaidya. Learning and incorporating shape models for semantic segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 203–211. Springer, 2017. 3

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 6

[34] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European conference on computer vision*, pages 347–365. Springer, 2020. 1, 2, 3

[35] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 3

[36] Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei. Superpixel-based segmentation for 3d prostate mr images. *IEEE transactions on medical imaging*, 35(3):791–801, 2015. 5, 6, 8

[37] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7158–7166, 2017. 2

[38] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[39] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1, 2, 3

[40] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018. 2, 3

[41] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 3

[42] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020. 3

[43] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 3, 7, 8

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2, 4, 6

[45] Clement Zotti, Zhiming Luo, Alain Lalande, and Pierre-Marc Jodoin. Convolutional neural network with shape prior applied to cardiac mri segmentation. *IEEE journal of biomedical and health informatics*, 23(3):1119–1128, 2018. 3