ACL-Net: Semi-supervised Polyp Segmentation via Affinity Contrastive Learning

Huisi Wu*, **Wende Xie, Jingyin Lin, Xinrong Guo**

College of Computer Science and Software Engineering, Shenzhen University
hswu@szu.edu.cn

Abstract

Automatic polyp segmentation from colonoscopy images is an essential prerequisite for the development of computer-assisted therapy. However, the complex semantic information and the blurred edges of polyps make segmentation extremely difficult. In this paper, we propose a novel semi-supervised polyp segmentation framework using affinity contrastive learning (ACL-Net), which is implemented between student and teacher networks to consistently refine the pseudo-labels for semi-supervised polyp segmentation. By aligning the affinity maps between the two branches, a better polyp region activation can be obtained to fully exploit the appearance-level context encoded in the feature maps, thereby improving the capability of capturing not only global localization and shape context, but also the local textural and boundary details. By utilizing the rich inter-image affinity context and establishing a global affinity context based on the memory bank, a cross-image affinity aggregation (CAA) module is also implemented to further refine the affinity aggregation between the two branches. By continuously and adaptively refining pseudo-labels with optimized affinity, we can improve the semi-supervised polyp segmentation based on the mutually reinforced knowledge interaction among contrastive learning and consistency learning iterations. Extensive experiments on five benchmark datasets, including Kvasir-SEG, CVC-ClinicDB, CVC-300, ColonDB and ETIS, demonstrate the effectiveness and superiority of our method. Codes are available at https://github.com/xiewende/ACL-Net.

Introduction

Colorectal cancer (CRC) is a common malignant tumor in the gastrointestinal tract and has become the third most common cancer in the world (Silva et al. 2014). Fortunately, CRC can be effectively prevented if polyps are removed in time. With the development of technology, automated segmentation of polyp plays a key role in the computer-aided diagnosis of CRC. Colonoscopy is an essential polyp detection method that can help doctors locate and remove polyps. However, precisely segmentation of polyps from colonoscopy videos is still a challenging task. First, the characteristics of polyps are highly variable, including various scales, locations, colors and textures (Figure 1(a)-(d)).

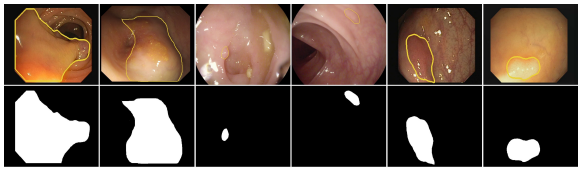


Figure 1: Challenges in semi-supervised polyp segmentation. (a)-(d) denote various scales, colors and textures of polyps. (c)-(f) illustrate the low contrast between polyps and surrounding tissues.

Second, the low contrast between the polyp and the background mucosa produces a blurred boundary (Figure 1 (e)-(f)), which may reduce the discrimination of object features and thus increase the possibility of incorrect segmentation.

Recently, supervised deep learning methods achieved remarkable success in polyp segmentation (Zhou et al. 2018; Fan et al. 2020; Kim, Lee, and Kim 2021; Wei et al. 2021). However, as the pixel-wise labeling process is particularly expensive and time consuming, the segmentation accuracy is still cannot guaranteed without enough high quality labelled datasets. To tackle this limitation, semi-supervised semantic segmentation methods are proposed to reduce the reliance on labels (Ke et al. 2020; Ouali, Hudelot, and Tami 2020).

Current state-of-the-art (SOTA) semi-supervised learning is mainly based on consistency regularization (Miyato et al. 2018; Mittal, Tatarchenko, and Brox 2019; Ouali, Hudelot, and Tami 2020; Liu et al. 2022; Seibold et al. 2022; Zhang et al. 2022) and contrastive learning (Zhao et al. 2022; Lai et al. 2021; Kwon and Kwak 2022; Wang et al. 2022b). Although these techniques can substantially improve the performance of semi-supervised semantic segmentation, they share a common drawback: weak learning ability at the initialization may carry bias due to less reliable pseudo-labels generated from the relatively poorer predictions of unlabeled images. Several existent methods are proposed to combine both contrastive learning and consistency learning, but they are usually optimized in different directions without knowledge transfer during the training process. However, different optimized directions may lead to a counterproductive effect. More recently, medical semi-supervised segmentation methods (Wu et al. 2021; Seibold et al. 2022) also simply

*Corresponding author. Email: hswu@szu.edu.cn

employed thresholding operations for different predictions to obtain pseudo-labels, which still cannot robustly identify low-contrast boundaries in polyps. Moreover, polyp semi-supervised segmentation (Wu et al. 2021) also adopted an adversarial learning strategy to obtain pseudo-labels, but the adversarial learning is also difficult to train and unstable. Different from the above methods, we propose an affinity contrastive learning (ACL) implemented between the student and teacher networks to consistently refine the pseudo-labels, which can provide a more reliable supervisory signal for unlabeled images during training.

In this paper, we propose a novel semi-supervised polyp segmentation framework based on affinity contrastive learning (ACL-Net), which is implemented between the student and teacher networks to consistently refine the pseudo-labels for semi-supervised polyp segmentation. By aligning the affinity maps between the two branches, we can obtain a better polyp region activation to fully exploit the appearance-level context encoded in the feature maps, thereby improving the capability of capturing not only global localization and shape context, but also the local textural and boundary details. To further refine the affinity aggregation between the two branches, we also implement a cross-image affinity aggregation (CAA) module to utilize the rich inter-image affinity context and establish a global affinity context based on the memory bank. Relying on the continuously and adaptively refining pseudo-labels based on the mutually reinforced knowledge interaction among contrastive learning and consistency learning iterations, we can finally improve the semi-supervised polyp segmentation with the optimized affinity maps. Extensive experiments on five polyp datasets, including Kvasir-SEG (Jha et al. 2020), CVC-ClinicDB (Bernal et al. 2015), CVC-300 (Vázquez et al. 2017), ColonDB (Bernal, Sánchez, and Vilarino 2012) and ETIS (Silva et al. 2014), have demonstrated the effectiveness and superiority of our proposed method. Our contributions are summarized as follows:

- We propose a novel semi-supervised polyp segmentation framework via affinity contrastive learning (ACL-Net), which can align the affinity maps generated in the student and teacher networks to more accurately capture the global appearance-level contexts.
- We propose a cross-image affinity aggregation (CAA) to enhance the knowledge transfer ability between contrasting learning and consistency learning iterations, thereby achieving a better refinement of pseudo labels.
- We demonstrate the effectiveness and advantages of our ACL-Net on five challenging polyp datasets, outperforming other competitors under different labelled conditions.

Related Work

Polyp Segmentation

Polyp detection and segmentation are effective techniques for computer-aided diagnosis, which can effectively prevent colorectal cancer. With the development of deep learning, automatic polyp segmentation based on convolutional neural networks (CNNs) has made extensively progressed in recent years (Zhou et al. 2018; Fan et al. 2020; Huang et al. 2021; Kim, Lee, and Kim 2021; Wei et al. 2021). However, most of them are based on fully supervised training strategies, which generally require large amounts of labeled data, and annotating the image is often labor-intensive and time-consuming. Therefore, semi-supervised polyp segmentation is a more promising approach to achieve satisfactory performance from limited labeling images.

Semi-supervised Semantic Segmentation

Previous semi-supervised methods (Hu et al. 2021a; Liu et al. 2022) were mainly based on pseudo-labels optimization and contrastive learning. Several methods (Lai et al. 2021; Kwon and Kwak 2022; Wang et al. 2022b) also demonstrated the superiority of employing both contrastive learning and self-training in the respective optimization at semi-supervised, without knowledge communication between them. However, semi-supervised methods for natural images usually still cannot model the overall appearance-level information of complex semantic information and low contrast of object regions in medical images. Differently, we propose a novel semi-supervised segmentation framework that combines both affinity contrasting learning and self-training learning to enhance the capability in capturing appearance-level context for semi-supervised segmentation.

Semi-supervised Medical Segmentation

RPG (Seibold et al. 2022) and BoostMIS (Zhang et al. 2022) proposed to define the pseudo-labels based on simply thresholding operations, which is still unreliable particularly for the polyp segmentation task with complex cases. CAFD (Wu et al. 2021) adopted two segmentation networks and discriminators to obtain higher confidence pseudo-labels, but the training of adversarial learning is still difficult and unstable. Different from the above approaches, we propose to utilize affinity contrastive learning to consistently learn high-quality appearance-level context to refine pseudo-labels, which also provide a better bridge for knowledge transfer between contrastive learning and consistency learning.

Method

Overview

The framework of our proposed ACL-Net is as illustrated in Figure 2, which is implemented under a famous mean-teacher architecture (Tarvainen and Valpola 2017a) for semi-supervised image segmentation. To fully exploit the appearance-level context maps, we propose an affinity contrastive learning (ACL) between the student and teacher networks to consistently refine the pseudo-labels for semi-supervised polyp segmentation. By aligning the affinity maps between the student and teacher networks, we can obtain a better polyp region activation to improve the capability of capturing not only global localization and shape context, but also the local textural and boundary details. Considering the rich inter-image affinity context among the unlabeled images, we also implement a cross-image affinity aggregation (CAA) module to further enhance the affinity aggregation between the student and teacher networks by establishing a global affinity context based on the memory bank. Finally,

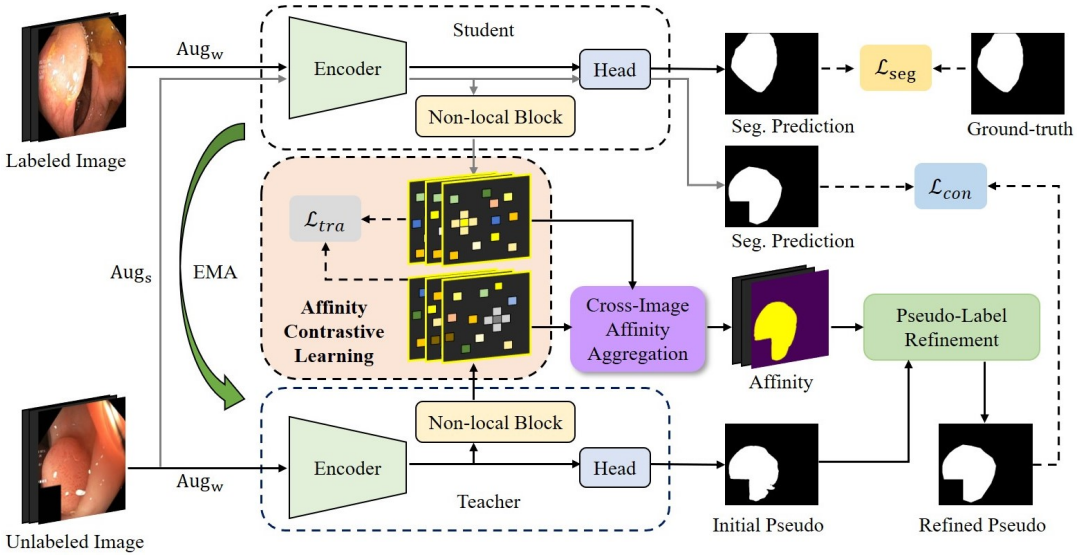


Figure 2: Overview of our proposed ACL-Net, where an affinity contrastive learning (ACL) mechanism is implemented between the student and teacher networks to consistently refine the pseudo-labels for semi-supervised polyp segmentation.

we can achieve a better semi-supervised polyp segmentation by continuously and adaptively refining the pseudo-labels based on the optimized affinity map.

Affinity Contrastive Learning

As shown in Figure 3, unlike existing fully supervised segmentation methods (Wang et al. 2022a; Ru et al. 2022), we propose an affinity contrastive learning (ACL) between student and teacher networks to consistently refine the pseudo-labels and enhance the semi-supervised polyp segmentation.

Given the feature maps F_s and F_t, which are extracted from the encoder of student network and teacher network respectively, we first apply the non-local self-attention block (Wang et al. 2018) on both F_s and F_t to the generate the affinity maps Aff_s and Aff_t. Considering that self-attention mechanism is essentially a directed graphical model (Veličković et al. 2017), the affinity matrix should be a symmetric structure. Therefore, we can obtain the affinity map by simply applying a 1×1 convolution layer to the feature map and its transpose, which can be denoted as:

$$\mathrm{Aff}_i = \mathrm{Conv}\left(\mathrm{NonLocal}\left(F_i\right) + \mathrm{NonLocal}\left(F_i\right)^T\right), \quad (1)$$

where $i \in (s, t)$ represents the student and teacher network respectively. T represents the matrix transpose operation. Conv is a 1×1 convolutional layer.

Since affinity reflects the correlations between a pixel point on the feature map and its neighbors, affinity maps Aff_s and Aff_t can capture more appearance-level context, including not only global localization and shape context, but also the local textural and boundary details. We next apply an element-wise summation between Aff_s and Aff_t to interact the appearance-level context of the different networks and enhance the reliability of the activation regions, described as $\mathrm{Aff}_m = \mathrm{Aff}_s \oplus \mathrm{Aff}_t$. Considering the relatively low contrast between polyps and their surrounding areas, it is a great challenge to segment the boundary areas accurately. In this regard, we can obtain the reliable foreground, background and uncertainty regions by simply filtering the affinity interaction map Aff_m based on two threshold values β_l and β_h, where $0 < \beta_l < \beta_h$. The filtered affinity interaction map $\mathrm{Aff}_{\hat{m}}$ can be written as:

$$\mathrm{Aff}_{\hat{m}}^{i,j} = \begin{cases} 1, & \text{if } \max\left(\mathrm{Aff}_m^{i,j,:}\right) \geq \beta_h, \\ 0, & \text{if } \max\left(\mathrm{Aff}_m^{i,j,:}\right) \leq \beta_l, \\ \mathrm{argmax}\left(\mathrm{Aff}_m^{i,j,:}\right), & \text{otherwise}, \end{cases} \quad (2)$$

where 0 and 1 denote the background and foreground, respectively. The $\mathrm{argmax}\left(\cdot\right)$ denotes to extract the semantic weight with the maximum uncertain regions.

So far, we can get affinity query $\mathrm{Aff}_q = \mathrm{Aff}_{\hat{m}} \otimes \mathrm{Aff}_s$ and affinity key $\mathrm{Aff}_k = \mathrm{Aff}_{\hat{m}} \otimes \mathrm{Aff}_t$ respectively, where $\otimes$ is matrix multiplication. Before calculating the contrastive loss, we need to calculate the foreground and background probabilities in the affinity interaction matrix, which can be obtained by filtering the uncertain regions, written as:

$$\tilde{y}_{\mathrm{Aff}_j} = \mathrm{argmax}\left(\mathrm{Aff}_j\right), \quad (3)$$

where $j \in (q, k)$ denotes affinity query and key respectively.

To align the affinity maps between the student and teacher networks, we need a contrastive loss $\mathcal{L}_{tra}$ to drive the positive pairs closer and push away the negative pairs. Meanwhile, we also adopt a memory bank $\mathcal{M}$ to store the features. Based on the standard form of contrastive loss defined in InfoNCE (Oord, Li, and Vinyals 2018), we can formulate the contrastive loss function for a query feature as:

$$\mathcal{L}_{tra} = -\log \frac{\mathrm{sim}\left(q, k_+\right)/\tau}{\mathrm{sim}\left(q, k_+\right)/\tau + \sum\limits_{k \in \mathcal{M}} \mathbb{I}_{q,k_-}\, \mathrm{sim}\left(q, k_-\right)/\tau}, \quad (4)$$

$$\mathrm{sim}(q, k) = \exp\left(\left(q \cdot k\right) / \left(\|q\|_2 \|k\|_2\right)\right), \quad (5)$$

$$\mathbb{I}_{q,k_-} = 1\left\{\tilde{y}_{\mathrm{Aff}_q} \neq \tilde{y}_{\mathrm{Aff}_k}\right\}, \quad (6)$$

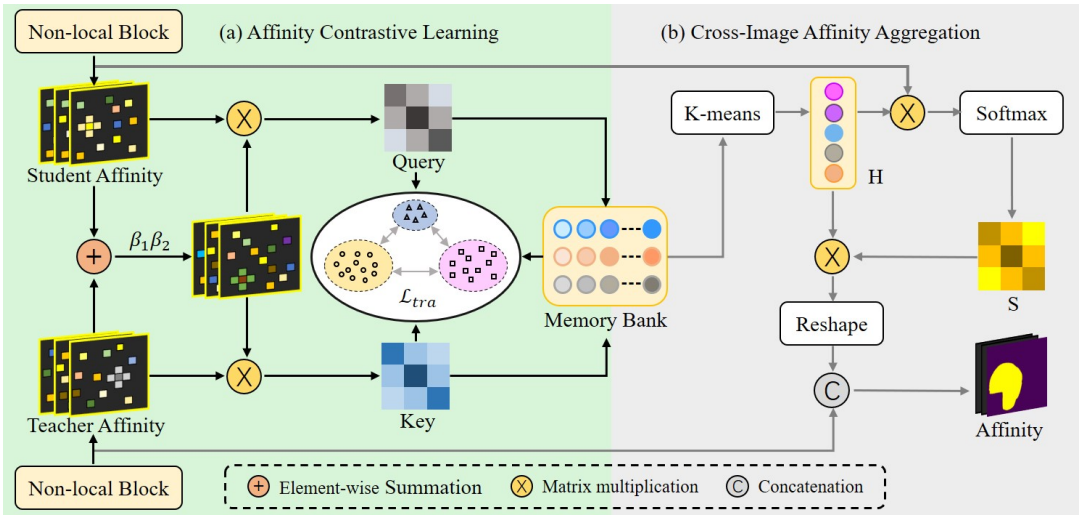


Figure 3: Illustration of Affinity Contrastive Learning (ACL) and Cross-Image Affinity Aggregation (CAA) modules.

where q is the anchor point from Aff_q. k_+ is a single positive, while the k_- includes all the negative samples. Note that the anchor q always locates in Aff_q, where its positive k_+ is on the same location in Aff_k. Negative samples k_- are obtained by filtering the features vectors in the memory bank according to the binary mask indicator Equation 6. Here τ indicates a temperature hyper-parameter. Moreover, we further implement a non-parameter and dynamic memory bank $\mathcal{M}$ (He et al. 2020; Wu et al. 2018; Xiao et al. 2017), which not only increase the negative samples to improve the contrastive learning optimization, but also provide a cross-image affinity to identify the global affinity context.

Cross-Image Affinity Aggregation

Considering that there has rich inter-image affinity context among the unlabeled images, we further propose a cross-image affinity aggregation (CAA) module to optimize the affinity between student and teacher networks. As shown in Figure 3, we first cluster the memory bank vectors into features H using the k-means, where k is experimentally set to 10 in our CAA module. In this regard, we can greatly reduce the dimension of the memory bank to achieve a much faster learning and inference speed. Then we can obtain an affinity matrix S bettwen the student affinity map Aff_s and the cross-image affinity context H, written as:

$$\mathrm{S} = \mathrm{Softmax}\left(\mathrm{Aff}_s \otimes H\right), \tag{7}$$

where $\mathrm{Softmax}$ denotes the normalization of each row for the input and $\otimes$ is matrix multiplication. Each element in S reflects the normalized similarity between each row in Aff_s and each column in H. Based on this affinity matrix S, we can further obtain a better cross-image affinity aggregation by the contextual summarization written as:

$$\mathrm{Aff}_{\mathrm{CAA}} = \mathrm{Concat}\left(\mathrm{Aff}_t, \mathrm{Reshape}\left(S \otimes H\right)\right), \tag{8}$$

where $\mathrm{Concat}\left(\cdot\right)$ denotes the concatenation operation. Obviously, Aff_{CAA} not only encodes intra-image local affinity context from both student and teacher networks, but also captures inter-image affinity context from other unlabeled images, thus achieving a better affinity aggregation.

Pseudo-Label Refinement

Considering that the initial pseudo-labels is still coarse and unreliable, we propose to adaptively refine the pseudo-labels with high- and low-level affinity. Inspired by (Ru et al. 2022), given the input unlabeled image x_u and the predictions $\hat{y}$ of teacher network, for the pixel at position (i, j) and (u, v), the channel and spatial pairwise terms can be defined as:

$$\kappa_{cha}^{ij,uv} = -\left(\frac{|I_{ij} - I_{uv}|}{w_1 \sigma_{cha}^{ij}}\right)^2, \quad \kappa_{spa}^{ij,uv} = -\left(\frac{|P_{ij} - P_{uv}|}{w_2 \sigma_{spa}^{ij}}\right)^2, \tag{9}$$

where I_{ij} and P_{ij} are the channel information and the spatial location of pixel (i, j), respectively. σ denotes the standard deviation and $w_{(1,2)}$ control the smoothness. Then, the affinity kernel can be constructed by the summation of κ_{cha} and κ_{spa}, which are normalized with a softmax, written as:

$$\kappa^{ij,uv} = \frac{\exp\left(\kappa_{cha}^{ij,uv}\right)}{\sum_{(x,y)} \exp\left(\kappa_{cha}^{ij,xy}\right)} + w_3 \frac{\exp\left(\kappa_{spa}^{ij,uv}\right)}{\sum_{(x,y)} \exp\left(\kappa_{spa}^{ij,xy}\right)}, \tag{10}$$

where (x, y) is sampled from the neighbor set of $\mathcal{H}(i, j)$. Inspired by the previous work (Araslanov and Roth 2020), we define $\mathcal{H}(i, j)$ as the 8-way neighbors with multiple dilation rates (1, 2, 4, 8, 12, and 24). Next, we can adaptively refine the initial pseudo-labels $\hat{y}$ from the teacher network according to low-level affinity kernel k and high-level affinity $\mathrm{Aff}_{\mathrm{CAA}}$ with multiple iterations, written as:

$$\hat{y}_t^{i,j} = \alpha_1 \sum_{(u,v)\in\mathcal{H}(i,j)} k^{ij,uv} \hat{y}_{t-1}^{u,v} + \alpha_2 \sum_{(u,v)\in\mathrm{Aff}_{\mathrm{CAA}}(i,j)} \mathrm{Aff}_{\mathrm{CAA}}^{ij,uv} \hat{y}_{t-1}^{u,v} \tag{11}$$

where t is a hyper-parameter to update iteration. α_1 and α_2 control weights of different affinities to refine pseudo-labels.

To provide the refined pseudo-labels for unlabeled images in the student network training, we can formulate the consistency loss as:

$$\mathcal{L}_{con} = \frac{1}{|\mathcal{N}_u|} \sum_{i=1}^{\mathcal{N}_u} \ell_{ce}\left(y_i^u, \hat{y}_i^u\right) + \ell_{dice}\left(y_i^u, \hat{y}_i^u\right), \tag{12}$$

where $\mathcal{N}_u$ is the number of unlabeled images. y_i^u is the prediction of the student network for $i\text{-}th$ unlabeled image and $\hat{y}_i^u$ represents the pseudo-labels optimized by Equation 11 in the teacher network for the $i\text{-}th$ unlabeled image.

Loss Functions

For the supervised branch, we exploited the labeled data to guide the network by minimizing the standard cross-entropy loss and dice loss (Soomro et al. 2018) in the student model with Equation 14. For the unsupervised branch, we first processed the images based on a weak-strong augmentation pair (Hu et al. 2021b), where several weak augmentations (image flipping, cropping and scaling) are applied to the input images of the teacher model, and strong augmentations (Chen et al. 2021; Ke et al. 2020) are also applied to the input images of the student model to improve the overall generalization. The contrastive loss $\mathcal{L}_{tra}$ in Equation 4 and the consistency loss $\mathcal{L}_{con}$ in Equation 12 can collaboratively drive the network to extract information from unlabeled images. The student model is optimized by minimizing the overall loss, which can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{tra}\mathcal{L}_{tra} + \lambda_{con}\mathcal{L}_{con}, \tag{13}$$

where λ_{tra} and λ_{con} are weights of corresponding loss components. Noted that the teacher model's weights are exponential moving average (EMA) updated by the student model's weights. The supervised loss $\mathcal{L}_{seg}$ is defined as:

$$\mathcal{L}_{seg} = \frac{1}{|\mathcal{N}_l|}\sum_{i=1}^{\mathcal{N}_l} \ell_{ce}\left(p_i^l, y_i^l\right) + \ell_{dice}\left(p_i^l, y_i^l\right), \tag{14}$$

where $\mathcal{N}_l$ is the number of labeled images. p_i^l represents the segmentation result of the student network for $i\text{-}th$ labeled image, while the y_i^l represents the corresponding label.

Experiments

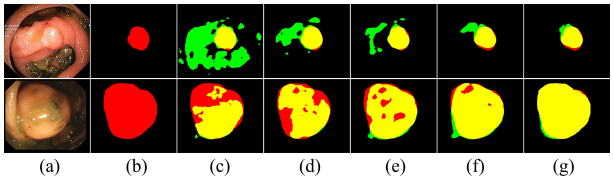
Datasets and Evaluation Metrics

We evaluated our method on five famous public polyp datasets, including Kvasir-SEG (Jha et al. 2020), CVC-ClinicDB (Bernal et al. 2015), CVC-300 (Vázquez et al. 2017), ColonDB (Bernal, Sánchez, and Vilarino 2012) and ETIS (Silva et al. 2014). Similar to the previous methods (Fan et al. 2020), a total of 1450 images, including 900 from Kvasir-SEG and 550 from CVC-ClinicDB, are divided into different labeled partition protocols (1/2, 1/4, 1/8) as our semi-supervised training datasets, and all above five datasets will be used in the inference phase.

In our experiments, we applied two widely-used metrics to evaluate the segmentation models, including mean dice (Dice (%)) and mean intersection over union (IoU (%)).

Implementation Details

Our proposed method is implemented with the PyTorch framework on a single NVDIA GeForce RTX 3090TI. ResNet-50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) and Transformer (Liu et al. 2021) are used as the backbone respectively, while DeepLabv3 (Chen et al. 2018) is selected as the segmentation head. The initial learning rate is set to 0.001, while the batch size is set to 8. We used a stochastic gradient descent (SGD) optimizer for training with a weight decay of 0.0001. We unified all images resolution to 384×384. To capture a priori knowledge of labeled images, we also performed 10 epochs of prewarm training before feeding the unlabeled images into the networks. We then adopt poly scheduling to schedule the learning rate, which is multiplied $\left(1 - \frac{iter}{total\_iter}\right)^{0.9}$. For the hyperparameter settings, both the loss function weights λ_{con} and λ_{tra} are experimentally set to 0.5. The weight of EMA is set to 0.999. We set the background scores $\beta_l = 0.45$ and $\beta_h = 0.75$ in Equation 2. The temperature parameter τ is 0.5 in Equation 4. In Equation 9 and Equation 10, we set the weight factors (w_1, w_2, w_3) as $(0.2, 0.2, 0.5)$ respectively. The weights (α_1,α_2) in Equation 11 are set to $(0.3, 0.7)$.



(a) (b) (c) (d) (e) (f) (g)

Figure 4: Visual comparisons of segmentation results extracted in ablation studies. (a) Input image. (b) Ground truth. (c) SupOnly. (d) With affinity contrastive learning. (e) and (f) are with low-level and high-level affinity refinement but without affinity aggregation, respectively. (g) Ours. Red, green and yellow regions represent the ground truth, prediction and their overlapping region respectively.

Method	TR	AC	AG	LA	HA	Dice	IoU
SupOnly						80.63	71.44
I	✓					81.29	73.67
II	✓	✓				83.92	75.86
III		✓		✓	✓	84.47	76.48
IV		✓	✓	✓		84.86	77.13
IV		✓	✓		✓	85.26	77.93
V		✓	✓	✓	✓	**85.89**	**78.51**

Table 1: Ablation studies of different components. TR: Threshold Refinement. AC: Affinity Contrast. AG: Affinity Aggregation. LA: Low-level Affinity Refinement. HA: High-level Affinity Refinement.

Ablation Studies

We performed extensive visual and statistical ablation experiments to verify the contribution of each component, as shown in Figure 4 and Table 1. The ablation experiments are conducted in the Kvasir-SEG by training the semi-supervised networks with different components on the 1/4 labeled images. We adopted ResNet-50 and Deeplabv3 as the backbone segmentation networks, and train the model only using labeled data, which is named as SupOnly method.

Ablation of Affinity Contrastive Learning. We optimized the affinity maps in student and teacher networks

Method	Kvasir-SEG						CVC-ClinicDB					
	1/2		1/4		1/8		1/2		1/4		1/8	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
SupOnly	82.21	74.19	80.63	71.44	79.54	69.31	80.26	72.48	78.42	70.88	77.05	68.71
MT	83.14	75.02	78.85	69.06	79.46	68.47	81.91	73.93	79.07	69.33	78.34	68.24
CAC	84.28	78.24	81.01	75.04	81.07	72.82	83.08	75.53	82.71	74.32	79.18	72.88
AEL	84.89	74.90	80.31	71.82	81.48	73.96	82.19	73.25	8042	71.88	80.87	72.47
CAFD	83.24	75.88	81.61	74.53	80.05	71.16	84.23	76.06	80.94	72.21	79.86	73.18
ELN	85.49	76.47	82.33	73.38	81.92	72.41	84.94	76.29	82.84	74.33	81.07	74.93
PSMT	85.17	77.58	83.18	75.33	82.31	72.82	84.51	77.58	81.52	73.96	80.71	74.24
U2PL	86.56	78.61	84.47	76.80	82.86	73.31	85.87	77.79	83.27	77.18	81.17	75.81
Ours(R)	**88.12**	**81.74**	**85.89**	**78.51**	**83.79**	**75.81**	**87.25**	**81.64**	**84.82**	**78.67**	**83.16**	**76.92**
Ours(T)	**86.89**	**80.94**	**85.09**	**77.83**	**82.66**	**75.26**	**87.72**	**82.14**	**84.53**	**78.74**	**82.36**	**76.21**

Table 2: Quantitative comparisons with different state-of-the-art methods on Kvasir-SEG and CVC-ClinicDB datasets.

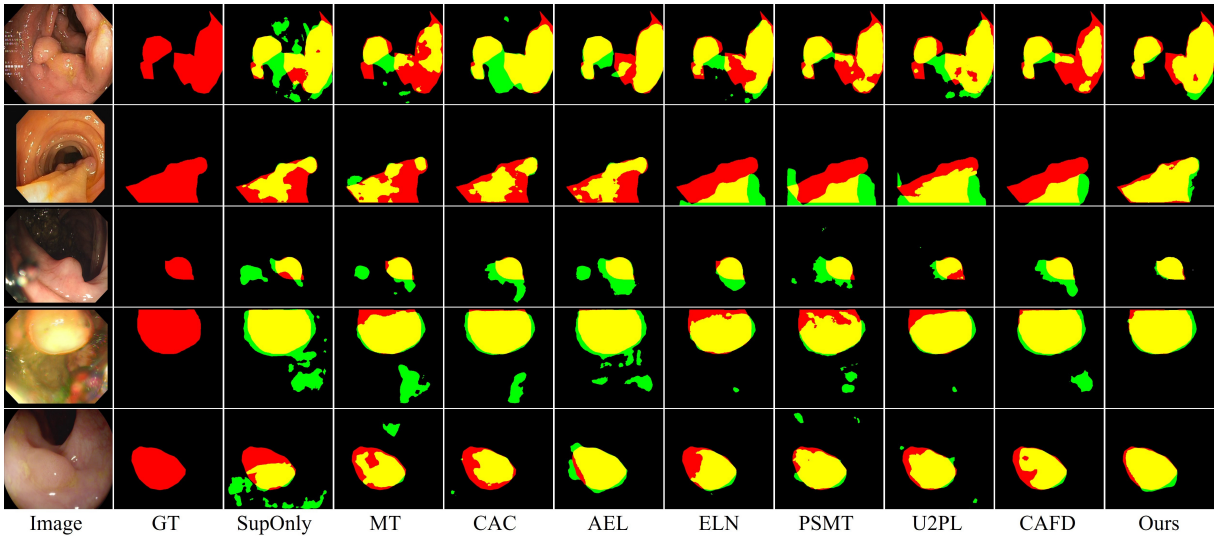


Figure 5: Visual comparisons with different state-of-the-art methods on the five public polyp datasets. The SupOnly method is trained on 1/4 labeled data only, while the other methods are trained on 1/4 labeled data and 3/4 unlabeled data. The images in the first to fifth rows are typically selected from Kvasir-SEG, CVC-ClinicDB, CVC-300, ColonDB and ETIS respectively. Red, green and yellow regions represent the ground truth, prediction and their overlapping region respectively.

by contrastive loss, to enhance the capability in capturing the rich appearance-level context. As shown in Table 1, by equipping with affinity contrastive learning module, Method II outperforms Method I in both Dice and IoU. Unlike Method I which only used thresholding to obtain the pseudo-labels, we can effectively extract invariant transferable appearance-level knowledge from labeled images to unlabelld images based on the affinity contrastive learning mechanism. Figure 6 also visually demonstrates the advantages of our affinity contrastive learning in feature space, where the T-SNE visualization shows a clear division between the background and the target regions.

Ablation of Cross-Image Affinity Aggregation. The comparison between Method III and Method V clearly demonstrates the effectiveness of cross-image affinity aggregation module. By fully exploiting the rich inter-image affinity context among the unlabeled images, we can further optimize the affinity aggregation between student and teacher networks, thereby improving the accuracy performance.

Ablation of Pseudo-Label Refinement. The comparisons among Method II, IV and V also can further verify the effectiveness of our pseudo-label refinement. By optimizing the initial pseudo-labels based on the strong supervised signal of affinity, we can obtain much better pseudo-labels with higher confidence to guide the semi-supervised learning via the consistency loss. With the pseudo-Label refinement module, we can obtain an improvement in Dice from 83.92% (Method II) to 85.89% (Method V).

Comparison with State-of-the-art Methods

We compared our approach with seven state-of-the-art methods: MT (Tarvainen and Valpola 2017b), CAC (Lai et al. 2021), AEL (Hu et al. 2021a), ELN (Kwon and Kwak 2022), PSMT (Liu et al. 2022), U2PL (Wang et al. 2022b),

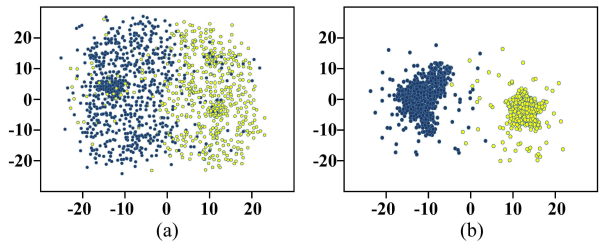


Figure 6: T-SNE visualization of features. (a) Backbone. (b) Affinity contrastive learning. The blue and yellow nodes represent background regions and target regions respectively.

Method	CVC-300		ColonDB		ETIS	
	Dice	IoU	Dice	IoU	Dice	IoU
SupOnly	79.09	69.47	58.86	48.97	49.21	41.55
MT	78.63	69.01	59.87	48.58	50.19	42.09
CAC	85.18	77.21	66.61	60.74	57.24	50.51
AEL	83.49	74.84	63.46	56.55	55.08	48.16
CAFD	82.21	74.08	61.78	53.94	54.26	48.19
ELN	83.86	75.22	62.34	55.73	55.93	47.77
PSMT	82.51	74.98	64.72	57.25	54.66	47.84
U2PL	84.36	76.06	65.05	57.12	56.27	48.85
Ours(R)	**86.47**	**79.54**	**68.71**	**62.65**	**58.89**	**52.44**
Ours(T)	**86.65**	**80.44**	**67.38**	**62.09**	**56.37**	**50.58**

Table 3: Generalization capability comparisons with different state-of-the-art methods. We first trained different networks by dividing the entire training image (1450) from Kvasir and CVC-ClinicDB into 1/4 labeled images (362) and 3/4 unlabeled images (1088). Statistical results are collected by directly applying the trained networks on CVC-300, ColonDB, and ETIS datasets.

and CAFD (Wu et al. 2021). We implemented all competitors through the same baseline segmentation network with ResNet-50 as the backbone as well as the same experimental environment and data augmentations to ensure fairness of the comparison. We define our method trained with ResNet-50 backbone as **Our(R)**, while with Transformer as **Our(T)**.

Learning Ability. Table 2 shows the advantages of our method in learning ability, where our approach is compared with other semi-supervised semantic segmentation approaches on the two visible datasets: Kvasir-SEG and CVC-ClinicDB datasets. We can clearly see that our method generally outperforms other competitors in different labeled partitions. Compared to the latest polyp semi-supervised method (CAFD), we obtain improvements of +4.28% in Dice and +4.01% in IoU using 1/4 labeled data.

Generalization Capability. We also conducted experiments to demonstrate the excellent generalization capability of our method on the three unseen datasets including CVC-300, ColonDB, and ETIS. As shown in Table 3, our approach also obtains a significant improvement over the other competitors, achieving 58.89% Dice and 52.44% IoU on the most challenging ETIS dataset. Relying on our novel affinity contrastive learning (ACL) mechanism implemented between student and teacher networks, we can consistently refine the pseudo-labels for semi-supervised polyp segmentation to obtain a better generalization capability.

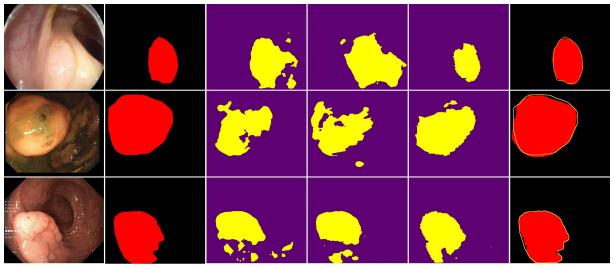


Figure 7: Visual comparisons of generated affinity maps. (a) Input image. (b) Ground truth. (c) Initial affinity map from student network. (d) Initial affinity map from teacher network. (e) Optimized affinity map with affinity contrast learning and cross-image aggregation. (f) Prediction.

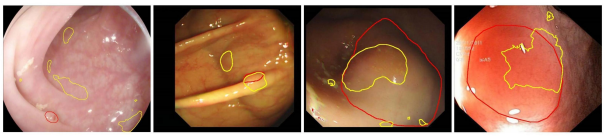


Figure 8: Failure cases. Yellow and red contours denote our segmented polyps and the ground truth, respectively.

Qualitative Results. In addition, we also performed a visual comparison among the SupOnly method, the seven SOTA methods, and our ACL-Net. As shown in Figure 5, we can also observe that our ACL-Net generally outperforms its competitors, especially for the challenging polyp cases with complex structures and small objects. Figure 7 further shows the optimization process of affinity map, demonstrating its excellent capability in extracting appearance-level context.

Limitations

Our method still has some limitations. Our method still may fail when the image contains multiple extremely small polyps (Figure 8 (a)-(b)), as well as when the color contrast between polyps and the background is extremely low (Figure 8 (c)-(d)).

Conclusion

In this work, we present a novel semi-supervised polyp segmentation framework via affinity contrastive learning (ACL-Net), which is implemented between the student and teacher networks to consistently refine the pseudo-labels. Specifically, we align the affinity maps between the two branches to obtain a better polyp region activation and fully exploit the appearance-level context. We also implement a cross-image affinity aggregation (CAA) module to utilize the rich inter-image affinity context and establish a global affinity context, thereby achieving a better semi-supervised polyp segmentation. Extensive experiments on five famous datasets demonstrate its effectiveness and superiorities.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 61973221 and 62273241), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011165), the COVID-19 Prevention Project of Guangdong Province, China (No. 2020KZDZX1174), the Major Project of the New Generation of Artificial Intelligence (No. 2018AAA0102900), and the Hong Kong Research Grant Council under General Research Fund Scheme (Project no. 15205919).

References

Araslanov, N.; and Roth, S. 2020. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4253–4262.

Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; and Vilariño, F. 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43: 99–111.

Bernal, J.; Sánchez, J.; and Vilarino, F. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9): 3166–3182.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, X.; Yuan, Y.; Zeng, G.; and Wang, J. 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273. Springer.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, H.; Wei, F.; Hu, H.; Ye, Q.; Cui, J.; and Wang, L. 2021a. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34: 22106–22118.

Hu, Z.; Yang, Z.; Hu, X.; and Nevatia, R. 2021b. Simple: similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15099–15108.

Huang, C.; Wu, H.; Lin, Y.; et al. 2021. A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean Dice and 86 FPS'. *arXiv preprint arXiv:2101.07172*.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; Lange, T. d.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, 451–462. Springer.

Ke, Z.; Qiu, D.; Li, K.; Yan, Q.; and Lau, R. W. 2020. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, 429–445. Springer.

Kim, T.; Lee, H.; and Kim, D. 2021. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2167–2175.

Kwon, D.; and Kwak, S. 2022. Semi-supervised Semantic Segmentation with Error Localization Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9957–9967.

Lai, X.; Tian, Z.; Jiang, L.; Liu, S.; Zhao, H.; Wang, L.; and Jia, J. 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1205–1214.

Liu, Y.; Tian, Y.; Chen, Y.; Liu, F.; Belagiannis, V.; and Carneiro, G. 2022. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4258–4267.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Mittal, S.; Tatarchenko, M.; and Brox, T. 2019. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4): 1369–1379.

Miyato, T.; Maeda, S.-i.; Koyama, M.; and Ishii, S. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8): 1979–1993.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Ru, L.; Zhan, Y.; Yu, B.; and Du, B. 2022. Learning Affinity from Attention: End-to-End Weakly-Supervised Semantic Segmentation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16846–16855.

Seibold, C. M.; Reiß, S.; Kleesiek, J.; and Stiefelhagen, R. 2022. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2171–2179.

Silva, J.; Histace, A.; Romain, O.; Dray, X.; and Granado, B. 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2): 283–293.

Soomro, T. A.; Afifi, A. J.; Gao, J.; Hellwich, O.; Paul, M.; and Zheng, L. 2018. Strided U-Net model: Retinal vessels segmentation using dice loss. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 1–8. IEEE.

Tarvainen, A.; and Valpola, H. 2017a. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tarvainen, A.; and Valpola, H. 2017b. Weight-averaged consistency targets improve semi-supervised deep learning results. CoRR abs/1703.01780. *arXiv preprint arXiv:1703.01780*, 1(5).

Vázquez, D.; Bernal, J.; Sánchez, F. J.; Fernández-Esparrach, G.; López, A. M.; Romero, A.; Drozdzal, M.; and Courville, A. 2017. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wang, W.; Feiszli, M.; Wang, H.; Malik, J.; and Tran, D. 2022a. Open-World Instance Segmentation: Exploiting Pseudo Ground Truth From Learned Pairwise Affinity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4422–4432.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Y.; Wang, H.; Shen, Y.; Fei, J.; Li, W.; Jin, G.; Wu, L.; Zhao, R.; and Le, X. 2022b. Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4248–4257.

Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S. K.; and Cui, S. 2021. Shallow attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 699–708. Springer.

Wu, H.; Chen, G.; Wen, Z.; and Qin, J. 2021. Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3489–3498.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xiao, T.; Li, S.; Wang, B.; Lin, L.; and Wang, X. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3415–3424.

Zhang, W.; Zhu, L.; Hallinan, J.; Zhang, S.; Makmur, A.; Cai, Q.; and Ooi, B. C. 2022. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20666–20676.

Zhao, X.; Fang, C.; Fan, D.-J.; Lin, X.; Gao, F.; and Li, G. 2022. Cross-Level Contrastive Learning and Consistency Constraint for Semi-Supervised Medical Image Segmentation. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 3–11. Springer.