Journal of Computers Vol. 35 No. 3, June 2024, pp. 363-377
doi: 10.53106/199115992024063503024

A Medical Image Segmentation Method Combining Knowledge Distillation and Contrastive Learning

Xiaoxuan Ma*, Sihan Shan, Dong Sui

School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China

maxiaoxuan@bucea.edu.cn, 2108550022051@stu.bucea.edu.cn, suidongcs2016@gmail.com

Received 1 May 2024; Revised 1 June 2024; Accepted 19 June 2024

Abstract. Recent advances in feature-based knowledge distillation have shown promise in computer vision, yet their direct application to medical image segmentation has been challenging due to the inherent high intra-class variance and class imbalance prevalent in medical images. This paper introduces a novel approach that synergizes knowledge distillation with contrastive learning to enhance the performance of student networks in medical image segmentation. By leveraging importance maps and region affinity graphs, our method encourages the student network to deeply explore the regional feature representations of the teacher network, capturing essential structural information and detailed features.This process is complemented by class-guided contrastive learning, which sharpens the discriminative capacity of the student network for different class features, specifically addressing intra-class variance and inter-class imbalance. Experimental validation on the colorectal cancer tumor dataset demonstrates notable improvements, with student networks ENet, MobileNetV2, and ResNet-18 achieving Dice coefficient score enhancements of 4.92%, 4.34%, and 4.59%, respectively. When benchmarked against teacher networks FANet, PSPNet, SwinUnet, and AttentionUnet, our best-performing student network exhibited performance boosts of 2.45%, 5.84%, 6.58%, and 3.56%, respectively, underscoring the efficacy of integrating knowledge distillation with contrastive learning for medical image segmentation.

Keywords: medical image segmentation, contrastive learning, knowledge distillation, lightweight network

1 Introduction

Medical image segmentation is the process of dividing the structures or tissues in medical images into different regions, to make it easier to analyze the images. All these segmentations help doctors and researchers to identify the location of the lesion, the severity of the lesion if any, the placement of the organ, etc. However, medical image segmentation is more difficult than semantic segmentation of natural scenes, with more complex lesion parts, varying sizes, and unknown unknowns. Deep learning technologies(especially Convolutional Neural Networks (CNNs)) have been applied to tackle these problems in medical image segmentation tasks. Moeskops et al. [1] first use this in early applications and developed a CNN for segmentation of six tissue classes in MRIs, pectoral muscles in MR breast images and coronary arteries. Alakwaa et al. [2] envisaged an automatic 3D-CNN-based framework to detect lung cancer by segmenting nodules in 3D CT scans. Vardhana et al. [3] applied hardware-accelerated CNNs to various lung CT images, brain MRI, X-ray images, and introduced CNNs to biomedical image segmentation. These studies provide examples of possible usage of deep learning in medical image segmentation and the robustness of deep learning in dealing with these challenging tasks.

The introduction of the U-Net architecture was a milestone for medical image segmentation, opening a series of creative attempts where eventually more dense connections, new feature extractors and applications of 3D convolutional kernels were used. This allows the RA-UNet [4] model, for instance, to combine an attention mechanism with the U-Net architecture [5]. U-Net++ [6]: A Nested U-Net Architecture (U-Net in U-Net) implemented a dense skip pathways through the layers of the encoder and decoder sub-networks to alleviate the semantic gap between the respective feature mappings. Other studies recognize that it is necessary to capture spatial continuity information, which directly increases the dimensionality of conventional kernels from 2D to 3D, like the networks 3DU-Net [7] and 3DU2-Net [8]. However, these methods come at the price of increasing computational costs and slower deployment. Therefore, to compensate for this, several researchers have studied

* Corresponding Author

lightweight networks, such as Enet [9] and ERFNet [10], which are suitable for real-time semantic segmentation, applications [11]. The problem is that we regularly simplify models to allow the model to train more quickly, and in doing so, we can often compromise the quality of the model itself.

As an approach to address the challenge of degrading model performance in lightweight networks, several techniques, such as model compression [12], transfer learning [13], and knowledge distillation (KD) [14], have been proposed in the literature. Of these, KD has been one of the most popular and successful methods in academia and industry. This improves the student model learning as it considers soft labels generated by a trained teacher model as an additional information to prevent overfitting [15]. In the literatures of the existing KD frameworks, Mean Teacher [16] is a popular forms and is being used in the field of medical image classification. A moving average over time updating the student model gives rise to feature distributions and predictions that are forced invariant to various perturbations, e.g. higher generalization ability with the small data, a stronger model sensible to noise. consequently, the secret to training highly accurate student networks is to distill knowledge of the teacher as much as possible. Relatively few works integrating efficiency studies with KD technology have focused on medical image segmentation problems in the past few years. Previous works have applied KD to chest X-rays [17] and 3D optical microscope images [18]. Most of these studies extract single sample-level knowledge from the teacher model, e.g. output logits [19] or feature maps [20]. Recently, Liu et al. [21] directed extraction toward specific small groups of samples The majority of current KD methods [20-25] were mainly borrowed from computer vision area and have not effectively addressed the following challenges in the medical domain. Medical datasets have larger intra-class variation and inter-class similarity, compared to data in the natural domain according to table_COUNTERS, row_HEP-2 Cell Strip Image Classification. In particular, two kinds of diseases could look so close to each other, in color, shape, texture, et. as opposed to two types of natural images (dogs vs cats case), overpowered softer attention mechanism for this task. Secondly, the medical image datasets tend to have severe class imbalance due to the nationwide prevalence and severity of some diseases (e.g.TB) or rarity of classes(e.g. ARDS). This means that what is being learned currently could be skewed towards the majority class and be less representative of the minority class.

In this paper, to solve the problem of high intraclass variance and the class imbalance in the medical images, we propose a method integrating the information distillation and contrastive learning, termed as Positive-Negative Contrastive Distillation Network (PNCD). This strategy tries to utilize the relevant information learned by sophisticated medical image segmentation networks for specialized purpose and adapt the established knowledge into a portable student network. The method mainly involves the following four modules:Region Matching Distillation (RMD), Region Affinity Distillation (RD), Positive-Negative Sample Contrastive Distillation (PNSCD), and Prediction Maps Distillation (PMD). These modules combine to make a good knowledge transfer from teacher network to student network.

The main idea of the RMD and RD modules is to provide a dedicated mechanism to transfer information in the intermediate layer importance maps and region affinity graphs from the teacher network to enhance the semantic understanding capability of the student network. The PNSCD module uses a class-guided contrastive manner to improve the feature discrimination of the student network. A PMD module then buzzes the student network to follow the final output of the teacher network, speeding up learning. Furthermore, it also gives a guarantee of the student network essentially by adding the loss of the segmentation task. To summarize, the major contributions of the paper are as follows:

(1) This paper introduces an architecture based on knowledge distillation, the Positive-Negative Contrastive Distillation Network (PNCD). It relies on four key modules: Region Matching Distillation (RMD), Region Affinity Distillation (RD), Positive-Negative Sample Contrastive Distillation (PNSCD), and Prediction Maps Distillation (PMD). By integrating knowledge distillation and contrastive learning, combined with the low training cost characteristics of lightweight models, it achieves relatively high accuracy in medical image segmentation.

(2) This paper designs a new Positive-Negative Sample Contrastive Distillation (PNSCD) module, which, by designing positive and negative samples for medical images, enables the student model's features to exhibit higher intra-class similarity and inter-class variance.

(3) This paper demonstrates the feasibility and reproducibility of the method by conducting robust experiments on a private medical image dataset and thoroughly considering ablation factors.

In this paper, we have organized the paper as follows: In Section 2: we discuss related work related to knowledge distillation and contrastive learning. Section 3: Principles of PNCD Structure Specifics To quantitatively evaluate the effectiveness of PNCD structure, we conduct comprehensive experiments in Section 4. Finally, Section 5 is used to conclude the paper.

2 Related Work

2.1 Knowledge Distillation

Knowledge Distillation (KD) [14] is a popular technique that enables complex and larger networks to train light-weight models to improve their performance without sacrificing on efficiency. As KD was first developed with the main idea of transferring the output, lately, there is a growing interest in the image segmentation community to distil feature and structural information as well. In this domain, He et al. [26] discussed a method to adapt knowledge from transferring long range dependencies, in semantic segmentation on natural images. It was the first successful demonstration that complex long-range relations exist as more than hand-crafted features in deep networks and can be exploited to improve segmentation tasks. Liu et al. [27] proposed a structured knowledge distillation to transfer pixel similarity to reinforce segmentation performance. Moreover, Xu et al. [28] investigated whether knowledges could be transferred by utilizing the different size of models for CT liver segmentation with a Growing Teacher Assistant Network (GTAN) strategy. Furthermore, Li et al. [29] introduced mutual KD to boost cross-modal segmentation as there are much discrepancies between CT and MRI, thereby demonstrating the necessity of translating prior knowledge to mitigate disparities between different imaging modalities. Qi et al. [30] proposed knowledge adaptation using the same teacher and student model networks and embedding the coordinate distillation that mixes channel and spatial features. Knowledge adaptation would enhance the transfer of information, thereby unfolding new perspectives on the brain segmentation tasks.

As the research continues, much more emphasis has been placed in this direction to find the best similarity between the features each of the teacher and the student model features. For example, He et al. [26] proposed an approach with an affinity distillation module which computes non-local interactions over the whole image to model long-range dependencies and then uses a distillation loss term to enforce the similarity between teacher and student features, thus facilitating the fusion between foreground and background. Liu et al. [31] proposed pixel-wise distillation and structured stacked distillation based on intermediate feature learning essentially two distillation schemes, pairwise distillation (pairwise-wise similarities are improved) and holistic (GANs that is used to improve global information). These two approaches complement each other and perform better than tradition model-based performance. Qin et al. [32] proposed a new module, importance mapping distillation, to rescale feature maps of a student to those of a teacher to match them even at the pixel level, thereby overcoming blurry boundaries in medical imaging and encoding detail internal to each semantic region for transfer. Ji et al. [33] aimed at transferring the structural and statistical knowledge of the texture from teacher networks to student networks by using contourlet decomposition and a novel texture intensity normalization module in early regions to allow the VGG-16 CNN model to capture more texture information and present it better. Wang et al. [34] introduced a technique to apply this idea on the missing modalities for medical imaging segmentation task and further in other domains using Dice Score as evaluation metric. Qin et al. [32] present an effective medical image segmentation method using knowledge distillation. It transfers some regional knowledge of teacher to the student, which speeds up the learning speed and improves the practical application efficiency. It provides a novel method for medical image analysis.

2.2 Contrastive Learning

Over the past couple of years, self-supervised learning, particularly in the to the point contrastive learning methods have made a lot of progress in computer vision applications and surpassed the performance of conventional supervised learning. Contrastive learning is used to learn representations with higher mutual information of different samples by pushing positives closer while pulling negatives further. This approach has been broadly utilized in different computer vision tasks [35-38], and a common objective of contrastive learning methods in the self-supervised training phase is to learn strong feature extractors that can be transferred later to downstream tasks [39-40]. For these methods, positive sample pairs are generated using data augmentation techniques [41-42], and a large number of negative sample pairs are essential for their success. Contrastive learning has been successful in the medical imaging realm. Chaitanya et al. [39] designed global and local contrastive losses based on structural similarities in 3D medical data, and utilized strategies rooted in the 3D volumetric medical image structural similarities to learn local areas of medical images, which offers a fresh perspective to choose positive and negative samples. Contrastive learning applied to Fully Supervised Semantic Segmentation. Wang et al. [43] presented a contrastive learning learned in pixel space, that effectively enhances the pixel-wise metric learning

by encouraging the pixel embeddings from the same class to be closer, compared to those from different classes. You et al. [44] presented an approach towards semi-supervised learning of anticorrelation in anatomical perception with contrastive distillation on medical images, updating an appropriate sample correction for the sample identification teaching set based on the anatomy or semantics for the better learning of the standard boundaries. Lee et al. [45] proposed a semantic-aware contrastive learning framework to learn the embedding that pulls two pixel embeddings of the same class closer and compute the similarity between two pixel embeddings more effectively, facilitating the multi-object segmentation on medical images. Li et al. [46] to solve the class imbalance problem in medical dataset a hierarchical instance comparison learning is presented. The technique addresses the issue of data imbalance and under-labeling in the data by learning from the majority class, in order to detect the minority class of diseases in data.

In this paper, we use contrastive loss to guide mutual learning between the student and teacher networks in the knowledge distillation model, aiming to maximize the student network's ability to learn from the teacher network.

3 Methodology

In this section, we will elaborate on the said method. The overall architecture designed in this paper is shown in Fig. 1. Blue structures correspond to the teacher network and yellow structures correspond to the student network. The input is a $W \times H$ grayscale medical image, and both teacher and student networks output a segmentation result in the same size. Four modules, i.e. Region Matching Distillation (RMD), Region Affinity Distillation (RD), Positive-Negative Sample Contrastive Distillation (PNSCD), and Prediction Maps Distillation (PMD) modules bridge the teacher and student structures with the knowledge distillation mechanism builtinto the method. RMD, RD, and PNSCD modules export intermediate information by building salient maps, region affinity graph, and creating class-guided contrastive formulation, respectively. The PMD Module forces the student network to replicate the teacher's final output, hastening the model's learning for segmentation. Subsequently, a loss specific to the segmentation task was introduced. This way, the student network can focus on segmenting the student network and the teacher network can extract experience. This will detail each module later on.

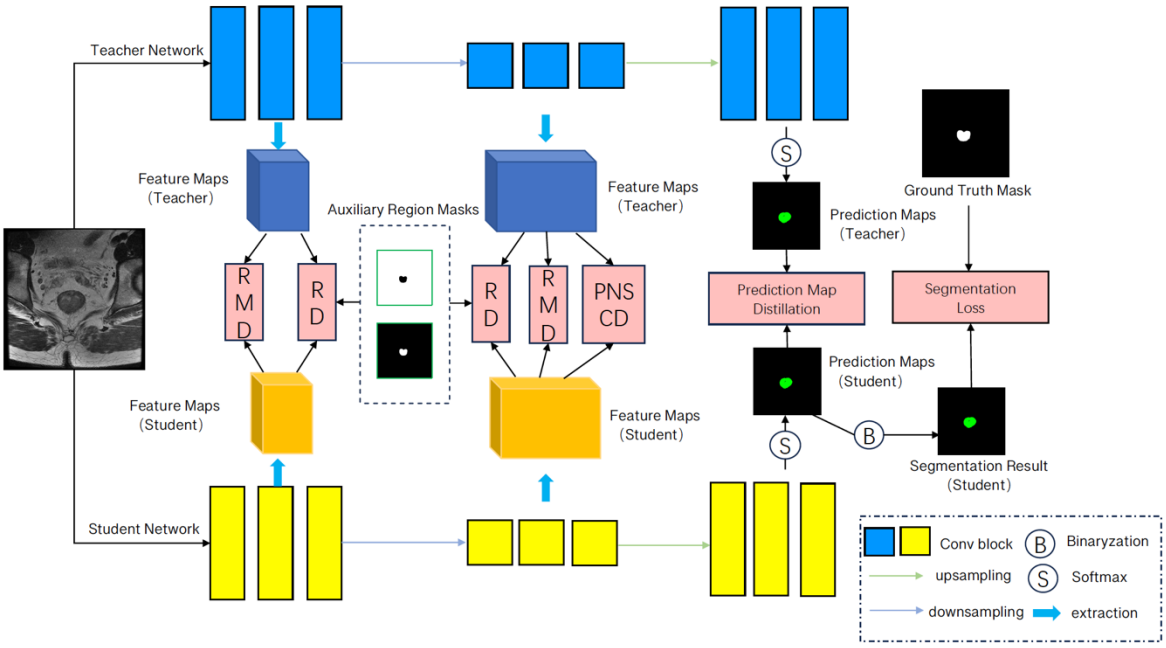

Fig. 1. The framework of Positive-Negative Contrastive Distillation Network (PNCD) module

3.1 Region Matching Distillation

Receiving knowledge from segmentation results is essential, but even more fundamental is the ability to learn segmentation process from the teacher network. The fundamental problem is that, two neural networks are not having same feature sizes, the feature sizes varies between the teacher and student networks. We therefore present an Region Matching Distillation (RMD) module for learning to transform the region matching across neural networks. The Detailed framework of this module is expressed in Fig. 2.

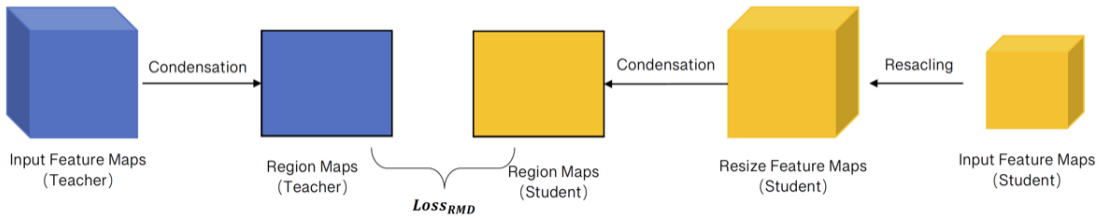

Fig. 2. The framework of Region Matching Distillation (RMD) module

More concretely, for a feature map E_s of size $c_s \times w_s \times h_s$ extracted from any layer of the student network and an associated feature map E_t of size $c_t \times w_t \times h_t$ extracted from the teacher network, we first perform a rescaling operation to make the spatial scales of the student and teacher feature maps E_t. This step can be defined as follows:

$$\widehat{e_s} = f\left(e_s\right); \widehat{e_s} \in R^{c_s \times w_t \times h_t}. \tag{1}$$

Whether unpooling or pooling are utilized depends on the spatial size relationship between E_s and E_t: respective smaller or larger or unaltered, the method $f(\cdot)$ selected.

For a feature map of size $C \times W \times H$, we create an importance aggregation map $M \in R^{w \times h}$ by summing the absolute value of ε along the channel dimension C. This process is defined as:

$$\varphi(\varepsilon) = \sum_{i=1}^{C} \left|\varepsilon_i\right|^2. \tag{2}$$

Where ε_i represents the ith matrix of ε along the channel dimension C.

Therefore, knowledge can be transferred by outputting its importance matching map, and the loss function for this module is as follows:

$$M_i^s = \varphi\left(f\left(e_i^s\right)\right); M_j^t = \varphi\left(e_j^t\right). \tag{3}$$

$$L_{RMD} = \sum_{(i,j) \in P} \left\| \frac{M_i^s}{\left\|M_i^s\right\|_2} - \frac{M_j^t}{\left\|M_j^t\right\|_2} \right\|_1. \tag{4}$$

e_i^s and e_j^t is the feature map extracted from the ith layer of student network and jth layer of teacher network, M_i^s and M_j^t is the importance matching map. P is the index pair set for all positions with the same embedding

size. $\|\cdot\|_1$ and $\|\cdot\|_2$ are the l_1 and l_2 norms, normalized.

3.2 Region Distillation

The crux is how to propagate the implicit structural information extracted by deep convolutional layers with large receptive field from teacher to student so as to enhance the student segmentation model performance even more. Motivated by this view, we introduce the Region Distillation (RD) module to transfer relational knowledge of areas from the teacher network to the student network.

We use labeled segmentation masks, including first-level feature maps extracted from specific regions of each type of semantic class Subsequently, we compute the similarity contrast values of regions for similarity between regional information. The structure of the RD module shown in Fig. 3. (Note that the RD module also take an extra input which is the auxiliary region mask in the Fig. 3.)

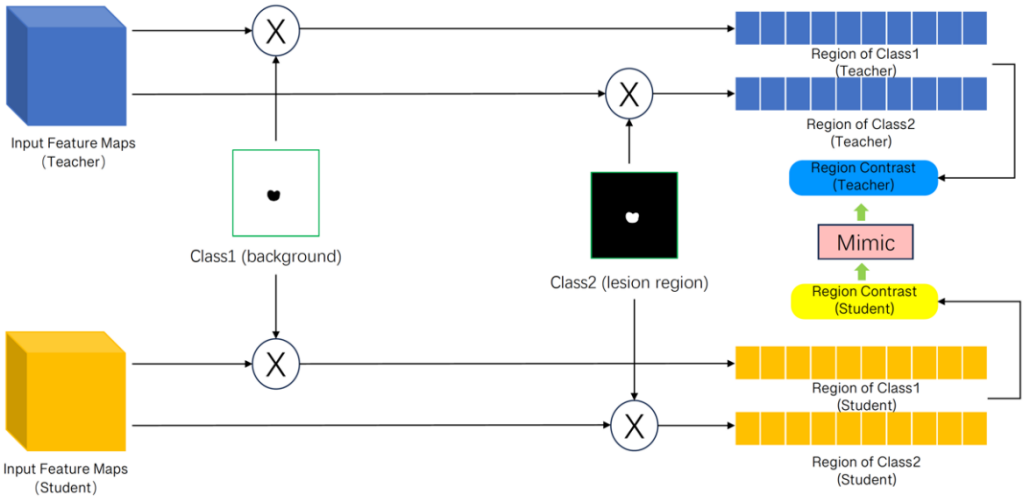

Fig. 3. The framework of Region Distillation (RD) module

That is more formally from a size feature map ε represented by, say, some intermediate layer, of size $C \times W \times H$. Because the feature map is not the same size as the input image, we need to resize the binary label mask from $W \times H$ to $w \times h$. Moreover, then for a semantic class i, we compute the region vector of R_i by averaging all features of length c in ε. Here is formula that can be used for this process:

$$R_i = \frac{1}{N_i} \sum_{j=1}^{w \times h} \varepsilon_j \cdot m_{ij}. \tag{5}$$

Where $i = 1, 2, \ldots, N_i$ is the ith pixel in the jth effective area of the ith mask. Finally, the regional contrast is obtained as:

$$V_{rc} = \frac{1}{n} \sum_{(i,j)} \frac{R_i^T R_j}{\|R_i\|_2 \|R_j\|_2}. \tag{6}$$

Where n represents the total number of possible class pairs. In the end, the loss for regional affinity is computed using the following loss function given the regional contrast values V_{rc}^s and V_{rc}^t of the student and teacher:

$$L_{RD} = \sum_{(i,j)\in P} \left\| V_{rc}^{s} - V_{rc}^{t} \right\|_{p}. \tag{7}$$

Where p is the type of norm. P denotes the index pairs set for positions sharing identical embedding sizes.

3.3 Positive-Negative Sample Contrastive Distillation

By incorporating contrastive learning into the traditional KD paradigm, Contrastive Representation Distillation (CRD) has obtained impressive distillation performance. To be specific, it obviously pulls away images between different classes, but it falsely separates images even from the same class to far distance in feature space and it grows the intra-class variance. To address this deficit, we propose a new Positive-Negative Sample Contrastive Distillation (PNSCD). In particular, PNSCD treats a pair of samples of the same class as a positive pair and moves the representations of them closer together, while a pair of image-level samples of different classes as negative pairs and separates the representations of them as illustrated in Fig. 4. Further, the feature embeddings of the teacher and the student are projected on the g_s and g_t through $Proj(\cdot)$, respectively, and g_s and g_t are normalized to the unit hypersphere via L_2 normalization to measure their similarity through the dot product. The loss for PNSCD is defined as:

$$L_{PNSCD} = -\frac{1}{k_p}\sum_{i=1}^{k_p}\left(\log\frac{e^{\left(g_s\cdot g_t^{\,i}/\tau\right)}}{e^{\left(g_s\cdot g_t^{\,i}/\tau\right)}+\frac{k_N}{M}} + \sum_{j=1}^{k_N}\log\left(1-\frac{e^{\left(g_s\cdot g_t^{\,i}/\tau\right)}}{e^{\left(g_s\cdot g_t^{\,i}/\tau\right)}+\frac{k_N}{M}}\right)\right). \tag{8}$$

Where τ is the temperature, which controls the concentration level of classes, L_{PNSCD}, M is the size of the dataset. In this case, L_{PNSCD} is minimized and therefore the student model is being trained to produce representations which are closer to the positive samples and further from the negative samples in the teacher model.

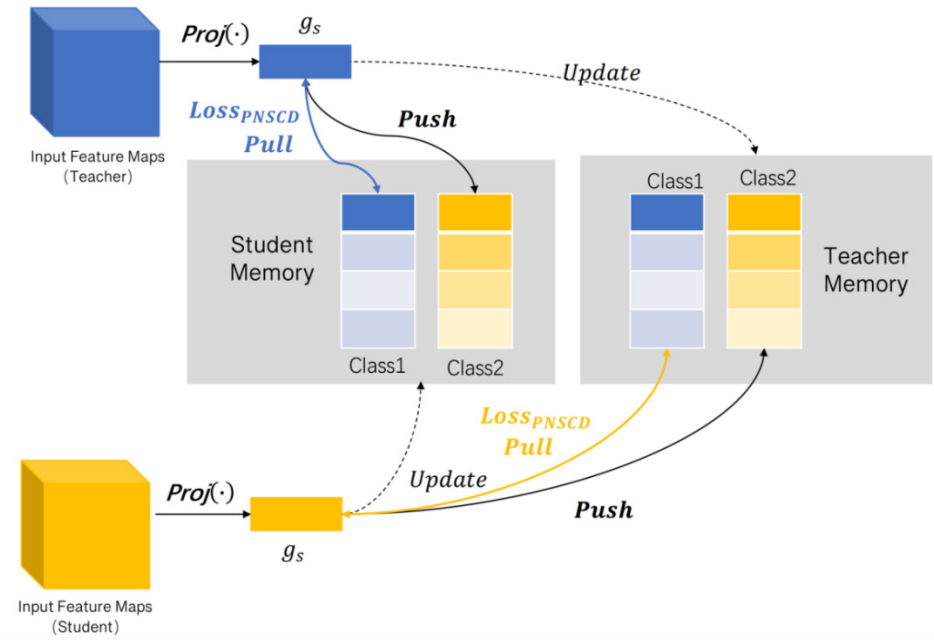

Fig. 4. The framework of Positive-Negative Sample Contrastive Distillation (PNSCD) module

3.4 Prediction Maps Distillation

Knowledge distillation [14] (Hinton et al., 2014) aims to make the student network replicate the teacher's final output, utilizing metrics such as cross-entropy and Kullback-Leibler (KL) divergence.In this paper, a Prediction Maps Distillation (PMD) module is introduced, and the main approach targets to do this by calculating the differences among the final layers. PMD further improves the teacher and student interrelationship by comparing the output segmentation maps of the two models to strengthen the student network from a spatial perspective. In this module, a pixel-aligned approach is employed, which compares the results on corresponding pixel locations of the teacher and student networks. Next, the Kullback-Leibler divergence function is used to measure the difference between the two to include all pixel dependencies. This computes the loss between all pixel pairs of the two networks at the same spatial location, and finally adds the loss to the total loss of the module. The loss function is given by the following:

$$L_{PMD} = \frac{1}{N} \sum_{i \in N} KL\left(p_i^s \| p_i^t\right). \tag{9}$$

Where, $N = W \times H$ calculates pixel count in a segmentation map. The Kullback-Leibler divergence function is $KL(\cdot)$. where p_i^s and p_i^t represent the probability for the *i*th pixel in the segmentation maps from the student and teacher networks respectively.

3.5 Training Process

Based on the PNCD method architecture diagram and combining the loss functions L_{RMD} , L_{RD} , L_{PNSCD} , and L_{PMD} , the total loss function $Loss_{total}$ for the PNCD is as follows:

$$Loss_{total} = Loss_{seg} + \alpha L_{PMD} + \beta_1 L_{RMD} + \beta_2 L_{RD} + \beta_3 L_{PNSCD}. \tag{10}$$

Where $Loss_{seg}$ is the general segmentation loss function, the hyperparameter α is set to 0.1, and β_1 , β_2 , and β_3 are all set to 0.9. In our experiments, it has been proven that the fluctuation of any single value is not sensitive. During the training process, we first utilize a pretrained teacher network, then minimize the total loss function $Loss_{total}$ to update the student network's parameters.This process includes not only the direct distillation comparing the outputs of the teacher and student networks but also involves the effective extraction and comparison of low-level and high-level features between them. Theoretically, we find that one can utilize any of the features of equal dimension for distillation, however, we also find that the selection of a few classes of low-level and high-level features that represent most of the feature space is more effective. It is noteworthy that the teacher network portion and the distillation modules are only used during the training phase and will be discarded during the testing phase. This means that the student network can predict independently after training is completed, without reliance on the teacher network or additional distillation mechanisms. This design not only ensures the lightweight and efficiency of the student network during inference but also demonstrates that carefully designed distillation strategies can significantly enhance the performance of the student network.

4 Experimental Results and Analysis

4.1 Setup

We experimented on modern segmentation architecture that mimicking a teacher network (DCSAU-Net) and a few lightweight networks (ENet) from an open-source community as a student network to show the efficiency of our distillation method. We can train these architectures separately for which we followed there official settings for network structures and hyperparameters. We optimized Adam with beta1 (0.9) and beta2 (0.999) for the distillation process and training all segmentation networks in experiments. Our initial learning rate was 0.001, and

we annealed the learning rate to 0.00001 using a cosine annealing scheduler. And data augmentation techniques such as random rotation and flipping. The actual MRI images that served as an input were 512 × 512 in dimensions, and another common pre-processing involved masking. The models to all the experiments are implemented with the PyTorch framework for training in a standard experimental setting. We failed to achieve all networks converge, so max trained 100 epochs.

4.2 Dataset

We gathered a novel dataset of 375 colorectal cancer tumors for this study, which included a series of MRI image datasets of colorectal cancer tumors from 2013 to 2020. This dataset was then split into training, testing and validation datasets that were used to train and evaluate the network. The main purpose to create this dataset is to collect the data for the region detection and segmentation of colorectal cancer. With this aim in mind, we have developed the methods presented in this paper and applied them to carry out these tasks before transitioning to clinical applications.

4.3 Evaluation Metrics

Dice index is widely used in evaluating medical image segmentation results. Since the Dice scores in our experiments are linked to segmentation tasks, it is important that methods are implemented in a practical and useful way. The per-case Dice metric function is defined as:

$$DICE(P,G) = \frac{2|P \cap G|}{|P| + |G|}. \tag{11}$$

Where P and G are the predicted and ground truth masks of the volumetric tumor, respectively.

In addition, two more metrics: Volume Overlap Error (VOE) and Relative Volume Difference (RVD) for volumetric overlap and annotation differences, other than Dice, which is the main segmentiation metric presented, were also introduced. Here is the functions of VOE and RVD.

$$VOE(P,G) = 1 - \frac{|P \cap G|}{|P| + |G|}. \tag{12}$$

$$RVD(P,G) = \frac{|P| - |G|}{|G|}. \tag{13}$$

4.4 Experimental Results

To validate that the proposed method is effective training and the trained method through distillation of number of different teacher and student networks as the first step. As teacher network structures, more complicated segmentation works, e.g. RA-UNet, DCSAU-Net, and UNet++ were used. In addition to the baseline and the existing light-weight models (ENet, MobileNet V2, ResNet-18), the student networks were some of the commonly used throughout those years. Following that, we conducted ablation experiments to validate each component and several largely adopted segmentation network structures which are employed as teacher networks in the distillation framework. Ideally, we compared them with the performances of the top student models as a benchmark for our method. Lastly, we will also look into the modules of our approach and the hyper parameters in the equations.

Main Experiments. After that, we use the distillation method proposed in this paper to dozens of teacher-student network pairs and verify it on our own dataset. The inconsistency of both number of upsampling and downsampling layers changes feature dimensions during the extraction of intermediate features. Therefore, we selected the first (low-level feature) and last (high-level feature) pairs embedding as well with the same sizes to feed as input in our distillation modules.

In this section of the experiment, we used the traditional medical segmentation model such as RA-UNet, a DCSAU-Net model, and the more relationship UNet++. Results are summarized in Table 1 for Dice, RVD, and VOE segmentation evaluation criteria (best results in bold), with the number of parameters in millions marked as M. The results indicated that with the help of teacher networks that have performed the best, all three student networks were subjected to significant improvements.

Table 1. Cross-experimental results on the PNCD method

Method	Dice	RVD	VOE	#Params (M)
Teacher				
T1: DCSAU-Net	0.8906	0.1690	-0.0019	6.3
T2: RA-Unet	0.8501	0.1714	0.0034	22.1
T3: Unet++	0.8182	0.1755	-0.0093	20.6
ENet	0.8074	0.1756	0.0104	0.353
ENet+T1 (ours)	**0.8566**	**0.1710**	**-0.0078**	
ENet+T2 (ours)	0.8270	0.1736	0.0120	
ENet+T3 (ours)	0.8125	0.1766	-0.0134	
ResNet18	0.7783	0.2243	0.0509	11.2
ResNet18+T1 (ours)	**0.8143**	0.1814	**0.0329**	
ResNet18+T2 (ours)	0.8031	0.1917	0.0388	
ResNet18+T3 (ours)	0.7884	**0.1804**	0.0427	
MobileNetV2	0.7810	0.1993	0.0398	2.2
MobileNetV2+T1 (ours)	**0.8344**	**0.1745**	**0.0154**	
MobileNetV2+T2 (ours)	0.8123	0.1896	0.0198	
MobileNetV2+T3 (ours)	0.8014	0.1842	0.0210	

As we can see in our results with several student networks trained, all of them benefited from their more powerful teacher networks, by employing our knowledge distillation method. Our approach is well-suited for medical image segmentation problems. The Dice coefficient scores for colorectal cancer tumor segmentation were improved by 4.92% (0.8566), 4.34% (0.8344) and 4.59% (0.8243) respectively over student networks, ENet, MobileNetV2, and ResNet-18. Among them, the ENet had the best improvement. In following experiments, ENet will be compared with other models.

Comparative Experiments. To demonstrate our model holistically we compare our learned student model against FANet, PSPNet, SwinUNet and AttentionUNet, which are classical network employed as teacher networks in the field of knowledge distillation. These two years ago we started with these selected networks, and they have been pioneers teacher networks doing knowledge distillation. Similarly, we picked DCSAU-Net as the teacher network for our chosen ENet network. Table 2 indicates experimental results with the best performance highlighted by a thick font style.

Table 2. Performance comparison between our method (PNCD) and current mainstream models used as teacher networks

Method	Dice	RVD	VOE
FANet	0.8321	0.1708	0.0082
PSPNet	0.7982	0.1755	-0.0093
SwinUnet	0.7908	0.1742	-0.0083
AttentionUnet	0.8210	0.1834	0.0110
ENet (ours)	**0.8566**	**0.1710**	**-0.0078**

This table clearly demonstrates that the Dice of the student model learned through our distillation framework is much higher than any of those of networks used as teacher models in other distillation frameworks. In particular, with comparing to FANet, PSPNet, SwinUnet and AttentionUnet, our student model achieved a performance increase of 2.45%, 5.84%, 6.58% and 3.56%, each. This result can better reflect the striking superiority of our proposed PNCD framework on our colorectal cancer tumor dataset built by us. Also, According to the visualisation results in Fig. 5. in the top row, since teacher network for the ENet network is DCSAU-Net, pixel-level segmentation map, we represent the background area in black and target area in green. Now, in Fig. 5. We can confirm ours being a superior method in terms of handling the colorectal cancer tumor dataset. By visualized cases, we find that in complex scenarios, our method can better approximate the ground truth segmentation and predict the lesion correctly when compared with other teacher models, which even outperforms other teacher models.

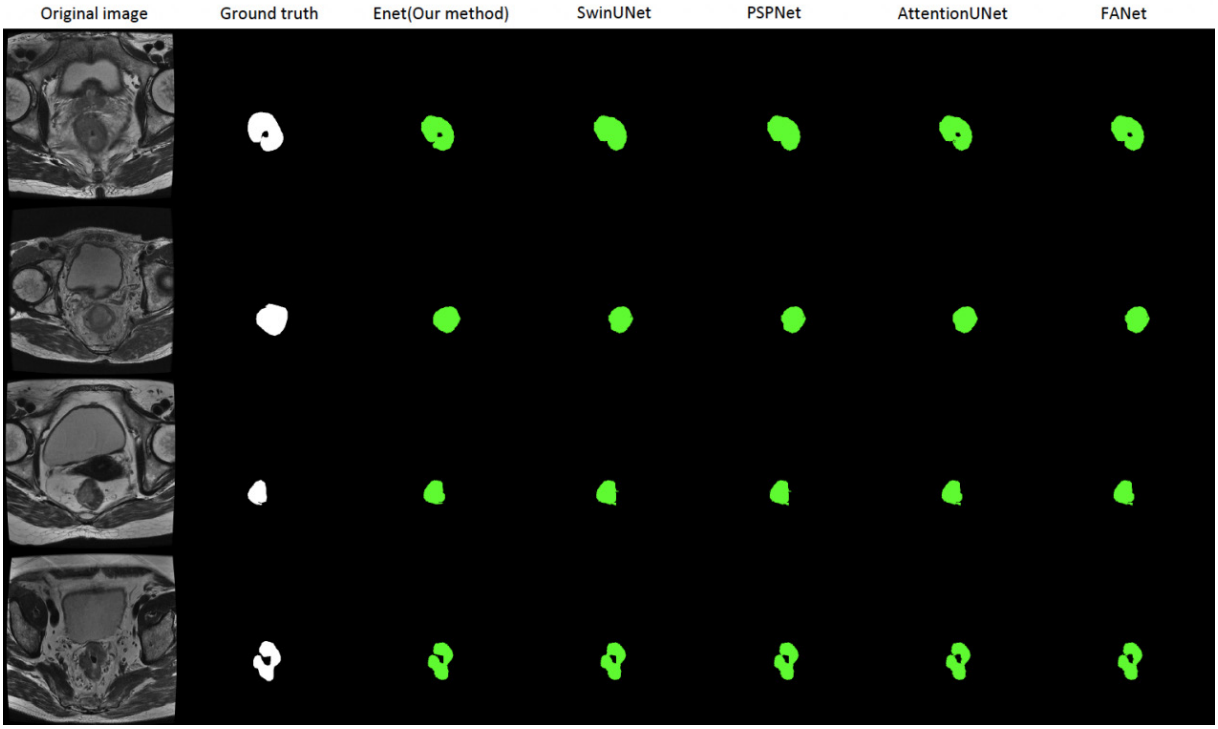

Fig. 5. Results of four sets of comparative experiments extracted from the dataset

Ablation Experiments. We performed the ablation experiments to carefully investigate the impact of each module in helping the student network better learn the knowledge from the tumor data of the colorectal cancer. The first three main contributions in this paper are the Region Matching Distillation (RMD) module, Region Distillation (RD) module, Positive-Negative Sample Contrastive Distillation (PNSCD) module, and Prediction Maps Distillation (PMD) module. Table 3: The Dice coefficient for colorectal cancer tumor, and for the this dataset other evaluation metrics, VOE and RVD Full size table Best performance results are highlight in bold. Experimental results reveal that the student network's performance significantly improved with the integration of any of these modules, as evidenced not only by a higher Dice coefficient but also by decreased VOE and RVD metrics. In addition, the experimental results showed that, when these modules were combined in our method, the optimal performance was achieved, which is illustrated in the last row of the experimental table.

In further analysis of the ablation experiments, we detailed the impact of each component on model performance. Specifically, we found that the Positive-Negative Sample Contrastive Distillation (PNSCD) module played a crucial role in enhancing the performance of the student network. Introducing the PNSCD module alone could increase the Dice coefficient score from 0.8074 to 0.8369, achieving a significant improvement of 2.69%. When we applied the Prediction Maps Distillation (PMD) module, Region Matching Distillation (RMD) module, Region Distillation (RD) module, and PNSCD module together in our distillation architecture, the model's Dice

coefficient score was further improved to 0.8566, a total increase of 4.92%. This result clearly demonstrates the positive role of each distillation component in enhancing model performance, especially the contribution of the PNSCD module. This not only confirms the effectiveness of our distillation architecture but also highlights the importance of considering different modules comprehensively when designing knowledge distillation strategies.

Table 3. Reliability of the components of our method on the dataset

Method	Dice	RVD	VOE
Teacher: DCSAU-Net	0.8906	0.1690	-0.0019
Student: Enet	0.8074	01756	0.0104
+PMD	0.8098	0.1820	-0.0112
+RMD	0.8133	0.1741	0.0113
+RD	0.8114	0.1810	0.0145
+PNSCD	0.8369	0.1698	0.0110
+PMD+RMD	0.8213	0.1723	0.0138
+PMD+RD	0.8310	0.1820	-0.0146
+PMD+PNSCD	0.8369	0.1698	-0.0110
+RMD+RD	0.8376	0.1789	0.0085
+RMD+PNSCD	0.8368	0.1745	-0.0094
+RD+PNSCD	0.8312	0.1723	0.0098
+PMD+RMD+RD	0.8387	0.1760	0.0121
+PMD+RD+PNSCD	0.8412	0.1892	0.0084
+RDM+RD+PNSCD	0.8436	0.1930	0.0113
+PMD+RMD+RD+PNSCD	**0.8566**	**0.1710**	**-0.0078**

Hyperparameter Insensitivity Experiments. In our study on hyperparameter sensitivity, we not only focused on the model's performance under the optimal values of given hyperparameters but also explored the impact on model performance after slight adjustments to these hyperparameters. Specifically, through experimentation, we found that the model achieved its best segmentation effect, 0.8566, when the optimal values of the four weight parameters in the given loss function, α, β_1, β_2, and β_3, were set to 0.1, 0.9, 0.9, and 0.9, respectively, as shown in Table 4. When these weights were set to 0, the training process was equivalent to training the original network, with bold indicating the best performance result.

Table 4. Experimental results for the component weights of hyperparameters α, β_1, β_2, and β_3 in $Loss_{seg}$

Method	Weight of components				Dice
	α	β_1	β_2	β_3	
Teacher: DCSAU-Net	0	0	0	0	0.8906
Student: ENet	0	0	0	0	0.8074
ENet + PNCD	0.1	0.9	0.9	0.9	0.8566
	0.2	0.9	0.9	0.9	0.8433
	0.1	1.8	1.8	0.9	0.8124
	0.1	1.8	0.9	1.8	0.8269
	0.1	0.9	1.8	1.8	0.8398

Notably, after doubling these weight values, we observed some relatively minor performance decreases. Even after doubling the four weight parameters α, β_1, β_2, and β_3, we could still notice a slight decrease in model performance. This suggests that while minor changes in hyperparameters might have some impact on the model's performance, our method is relatively insensitive to the selection of hyperparameters.

Based on these findings, we are inclined to use the best values for hyperparameters in practical applications to ensure that the model can achieve optimal performance. Although we also recognize that there may be some flexibility in the choice of hyperparameters, for stable and reliable model performance, we recommend prioritizing the optimal hyperparameter values previously proven by experiments in practice. Doing so not only simplifies the hyperparameter tuning process but also enhances the reproducibility and stability of the model.

5 Conclusion

This paper proposes a distillation approach that combines knowledge distillation and contrastive learning, which can empower lightweight student networks in mastering formidable medical image segmentation missions. Our method, comprising four different novel modules: Region Matching Distillation (RMD), Region Affinity Distillation (RD), Positive-Negative Sample Contrastive Distillation (PNSCD), and Prediction Maps Distillation (PMD), allows for the deep and efficient transfer of knowledge. The deep intermediate layer information of the teacher network is taught to the student network by the student network, not only deepening the understanding of the student network about the information of the middle layer of the teacher network, but also improving the segmentation accuracy of the student network by means of precise prediction mapping and contrastive learning. Furthermore, our approach unlocks a new solution for computationally restricted setting on medical image segmentation, and illuminates the way on combining KD and CL in a broader range of medical image analysis areas.

While our results on specific datasets affirm the effectiveness of the proposed method, the challenges in medical image segmentation are diverse, encompassing various disease types and imaging techniques. Therefore, future efforts will focus on assessing the performance of our approach across different medical imaging datasets to ensure its broad applicability and effectiveness.

6 Acknowledgement

This research was funded by National Key Research and Development Program of China [grant number 2020YFB2103604].

References

[1] P. Moeskops, J.M. Wolterink, B.H. Van Der Velden, K.G. Gilhuijs, T. Leiner, M.A. Viergever, I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in: Proc. 2016 Medical Image Computing and Computer-Assisted Intervention, 2016.

[2] W. Alakwaa, M. Nassef, A. Badr, Lung cancer detection and classification with 3D convolutional neural network (3D-CNN), International Journal of Advanced Computer Science and Applications 8(8)(2017) 409-417.

[3] M. Vardhana, N. Arunkumar, S. Lasrado, E. Abdulhay, G. Ramirez-Gonzalez, Convolutional neural network for bio-medical image segmentation with hardware acceleration, Cognitive Systems Research 50(2018) 10-14.

[4] Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans, Frontiers in Bioengineering and Biotechnology 8(2020) 605132.

[5] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. 2015 Medical Image Computing and Computer-Assisted Intervention, 2015.

[6] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Proc. 2018 Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 2018.

[7] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: Proc. 2016 Medical Image Computing and Computer-Assisted Intervention, 2016.

[8] C. Huang, H. Han, Q. Yao, S. Zhu, S.K. Zhou, 3D U 2-Net: A 3D universal U-Net for multi-domain medical image segmentation, in: Proc. 2019 International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019.

[9] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation. <https://arxiv.org/abs/1606.02147>, 2016 (accessed 28.12.2023)

[10] E. Romera, J.M. Alvarez, L.M. Bergasa, R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Transactions on Intelligent Transportation Systems 19(1)(2017) 263-272.

[11] D. Jha, S. Ali, N.K. Tomar, H.D. Johansen, D. Johansen, J. Rittscher, M.A. Riegler, P. Halvorsen, Real-time polyp detection, localization and segmentation in colonoscopy using deep learning, IEEE Access 9(2021) 40496-40510.

[12] C. Buciluǎ, R. Caruana, A. Niculescu-Mizil, Model compression, in: Proc. 2006 12th ACM SIGKDD International Conference on Knowledge discovery and data mining, 2006.

[13] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big data 3(2016) 1-40.

[14] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network. <https://arxiv.org/abs/1503.02531>, 2014 (accessed 03.01.2023)

[15] P. Chen, S. Liu, H. Zhao, J. Jia, Distilling knowledge via knowledge review, in: Proc. 2021 IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2021.
[16] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Proc. 2017 31st Conference on Neural Information Processing Systems, 2017.
[17] T.K.K. Ho, J. Gwak, Utilizing knowledge distillation in deep learning for classification of chest X-ray abnormalities, IEEE Access 8(2020) 160749-160761.
[18] H. Wang, D. Zhang, Y. Song, S. Liu, Y. Wang, D. Feng, H. Peng, W. Cai, Segmenting neuronal structure in 3D optical microscope images via knowledge distillation with teacher-student network, in: Proc. 2019 IEEE 16th International Symposium on Biomedical Imaging, 2019.
[19] J.J. Thiagarajan, S. Kashyap, A. Karargyris, Distill-to-label: weakly supervised instance labeling using knowledge distillation, in: Proc. 2019 18th IEEE International Conference On Machine Learning And Applications, 2019.
[20] J. Wu, S. Yu, W. Chen, K. Ma, R. Fu, H. Liu, X. Di, Y. Zheng, Leveraging undiagnosed data for glaucoma classification with teacher-student learning, in: Proc. 2020 Medical Image Computing and Computer Assisted Intervention, 2020.
[21] Q. Liu, L. Yu, L. Luo, Q. Dou, P.A. Heng, Semi-supervised medical image classification with relation-driven self-ensembling model, IEEE transactions on medical imaging 39(11)(2020) 3429-3440.
[22] B. Unnikrishnan, C.M. Nguyen, S. Balaram, C.S. Foo, P. Krishnaswamy, Semi-supervised classification of diagnostic radiographs with noteacher: A teacher that is not mean, in: Proc. 2020 Medical Image Computing and Computer Assisted Intervention, 2020.
[23] S. Abbasi, M. Hajabdollahi, P. Khadivi, N. Karimi, R. Roshandel, S. Shirani, S. Samavi, Classification of diabetic retinopathy using unlabeled data and knowledge distillation, Artificial Intelligence in Medicine 121(2021) 102176.
[24] A. Patra, Y. Cai, P. Chatelain, H. Sharma, L. Drukker, A.T. Papageorghiou, J.A. Noble, Efficient ultrasound image analysis models with sonographer gaze assisted distillatio, in: Proc. 2019 Medical Image Computing and Computer Assisted Intervention, 2019.
[25] Y. Hou, Z. Ma, C. Liu, C.C. Loy, Learning lightweight lane detection cnns by self attention distillation, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[26] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, in: Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[27] Y. Liu, C. Shu, J. Wang, C. Shen, Structured knowledge distillation for dense prediction, IEEE transactions on pattern analysis and machine intelligence 45(6)(2020) 7035-7049.
[28] P. Xu, K. Kim, J. Koh, D. Wu, Y.R. Lee, S.Y. Park, W.T. Tak, H. Liu, Q. Li, Efficient knowledge distillation for liver CT segmentation using growing assistant network, Physics in Medicine & Biology 66(23)(2021) 235005.
[29] K. Li, L. Yu, S. Wang, P.A. Heng, Towards cross-modality medical image segmentation with online mutual knowledge distillation, in: Proc. 2020 AAAI Conference on Artificial Intelligence, 2020.
[30] Y. Qi, W. Zhang, X. Wang, X. You, S. Hu, J. Chen, Efficient knowledge distillation for brain tumor segmentation, Applied Sciences 12(23)(2022) 11980.
[31] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, in: Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[32] D. Qin, J.J. Bu, Z. Liu, X. Shen, S. Zhou, J.J. Gu, Z.H. Wang, L. Wu, H.F. Dai, Efficient medical image segmentation based on knowledge distillation, IEEE Transactions on Medical Imaging 40(12)(2021) 3820-3831.
[33] D. Ji, H. Wang, M. Tao, J. Huang, X.S. Hua, H. Lu, Structural and statistical texture knowledge distillation for semantic segmentation, in: Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[34] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, R. Li, Prototype knowledge distillation for medical segmentation with missing modality, in: Proc. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, 2023.
[35] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proc. 2020 International Conference on Machine Learning, 2020.
[36] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, in: Proc. 2020 34th Conference on Neural Information Processing Systems, 2020.
[37] I. Misra, L.V.D. Maaten, Self-supervised learning of pretext-invariant representations, in: Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[38] J. Peng, P. Wang, C. Desrosiers, M. Pedersoli, Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels, Advances in Neural Information Processing Systems 34(2021) 16686-16699.
[39] K. Chaitanya, E. Erdil, N. Karani, E. Konukoglu, Contrastive learning of global and local features for medical image segmentation with limited annotations, in: Proc. 2020 34th Conference on Neural Information Processing Systems, 2020.
[40] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, Y. Wang, Semi-supervised medical image segmentation via a tripled uncertainty guided mean teacher model with contrastive learning, Medical Image Analysis 79(2022) 102447.
[41] Y.M. Asano, C. Rupprecht, A. Vedaldi, A critical analysis of self-supervision, or what we can learn from a single image. <https://arxiv.org/abs/1904.13132>, 2020 (accessed 26.01.2023)
[42] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Proc. 2022 Machine Learning for Healthcare Conference, 2022.
[43] W. Wang, T. Zhou, F. Yu, J. Daim, E. Konukoglu, L. Van Gool, Exploring cross-image pixel contrast for semantic segmentation, in: Proc. 2021 IEEE/CVF International Conference on Computer Vision, 2021.

[44] C. You, W. Dai, Y. Min, L. Staib, J.S. Duncan, Bootstrapping semi-supervised medical image segmentation with anatomical-aware contrastive distillation, in: Proc. 2023 International Conference on Information processing in Medical Imaging, 2023.
[45] H.H. Lee, T. Tang, Q. Yang, X. Yu, L.Y. Cai, L.W. Remediios, S. Bao, B.A. Landman, Y. Huo, Semantic-Aware Contrastive Learning for Multi-object Medical Image Segmentation, IEEE Journal of Biomedical and Health Informatics 27(9)(2023) 4444-4453.
[46] Y. Li, G. Qian, X. Jiang, Z. Jiang, S. Zhang, K. Li, Q. Lao, Hierarchical-instance contrastive learning for minority detection on imbalanced medical datasets, IEEE Transactions on Medical Imaging 43(2023) 416-426.