CoCo DistillNet: a Cross-layer Correlation Distillation Network for Pathological Gastric Cancer Segmentation

Wenxuan Zou[1], Muyi Sun[2,†]

[1]School of Automation, Beijing University of Posts and Telecommunications, Beijing, China
[2]Center for Research on Intelligent Perception and Computing, NLPR, CASIA, Beijing, China

zouwenxuan@bupt.edu.cn, muyi.sun@cripac.ia.ac.cn

Abstract—In recent years, deep convolutional neural networks have made significant advances in pathology image segmentation. However, pathology image segmentation encounters with a dilemma in which the higher-performance networks generally require more computational resources and storage. This phenomenon limits the employment of high-accuracy networks in real scenes due to the inherent high-resolution of pathological images. To tackle this problem, we propose CoCo DistillNet, a novel Cross-layer Correlation (CoCo) knowledge distillation network for pathological gastric cancer segmentation. Knowledge distillation, a general technique which aims at improving the performance of a compact network through knowledge transfer from a cumbersome network. Concretely, our CoCo DistillNet models the correlations of channel-mixed spatial similarity between different layers and then transfers this knowledge from a pre-trained cumbersome teacher network to a non-trained compact student network. In addition, we also utilize the adversarial learning strategy to further prompt the distilling procedure which is called Adversarial Distillation (AD). Furthermore, to stabilize our training procedure, we make the use of the unsupervised Paraphraser Module (PM) to boost the knowledge paraphrase in the teacher network. As a result, extensive experiments conducted on the Gastric Cancer Segmentation Dataset demonstrate the prominent ability of CoCo DistillNet which achieves state-of-the-art performance.

Index Terms—knowledge distillation, model compression, gastric cancer segmentation

I. INTRODUCTION

In clinic medicine, pathology image segmentation serves as a beneficial measure to help pathologists diagnosis the region of cancerization, tumors and so on. Recently, with the development of deep learning, various delicate Convolution Neural Networks (CNNs) are devised and applied in the pathology image segmentation. Simultaneously, such intricate networks exhibit state-of-the-art performance in the area of automatic pathology image segmentation [1], [2].

Nevertheless, high-performance networks usually involve complicated, deep, and wide network architecture with numerous parameters. Furthermore, the size of Whole-Side Images (WSI) (e.g. 100,000×50,000 pixels) which are commonly used in digital pathology is over 100 times a natural image [2]. However, such high-resolution images cannot be sent into

† Corresponding authors.

CNNs as inputs directly. To this end, previous researchers take the patch-based strategy as regular approach [2] to handle this issues. Nevertheless, this strategy aggravates the requirement of memory and augments inference time. As drawbacks aforementioned, training an appropriate network with high-performance and low consumption, as well as loading the model into embedded devices is nearly impractical, taking account of limited computation resources and storage of the medical devices in real scenes.

There have been various endeavors which engage in conquering the computation limitation and memory constraint in portable or edge devices. Model compression is one of the main directions which could be roughy categorized into three classes: network quantization [4], [5], network pruning [6], [7], and knowledge distillation [10]–[13]. The purpose of network quantization is to transform the weights and features into low bitwidth integer types. Network pruning can compress the size of networks by cutting off some non-essential weights or neurons. Knowledge distillation is conducted to transfer the knowledge of a pre-trained cumbersome teacher network to a compact student network in which the knowledge carries more inter-class and intra-class information than labels. However, network quantization exhibits poor performance in the large CNNs [8], while the network pruning criteria demands hand-craft sensitivity for layers [9]. Furthermore, it is worth noting that network pruning or quantization need specific devices or implementations for accelerating CNNs. On the contrary, knowledge distillation can be applied in various network architectures (e.g. collaboration with lightweight CNNs). Meanwhile, this method appears high efficiency with high accuracy and low consumption, as the student network can acquire knowledge from the teacher network that significantly outperforms the student. In addition, knowledge distillation does not need extra parameters whether in the process of training or inference.

In this paper, we adopt knowledge distillation to circumvent the tradeoff between accuracy and efficiency in pathology segmentation. Initially, knowledge distillation is defined as a method that a student network mimics the 'soft targets' from a teacher network, i.e. the prediction of teacher network can supply extra supervision to train the student network [10].

Next, knowledge distillation is extended to approximate the middle features of CNNs [11]. However, different network architectures are prone to extract different features in the middle layer of networks. Consequently, directly mimicking the middle features between student and teacher is inappropriate when student network has different architecture from the teacher network [12]. Subsequently, many researchers study the similarity of inter-class and intra-class in the image classification task [12], [14], [15]. Considering the structural property of image segmentation, i.e. the relations among pixels in a feature map, researchers tend to transfer pixel-wise similarity of the teacher network to guide the training of the student network [13], [16]–[18].

The state-of-the-art networks have a common property, which tend to extract low-level representations in shallow layers as well as high-level representations in depth layers. However, recent knowledge distillation methods for image segmentation pay less attention to the correlation between low-level and high-level representations. Motivated by the Flow of Solution Procedure (FSP matrix) [19], we propose a novel distillation method that defines Cross-layer Correlations (CoCo) as transferred knowledge. However, direct computation of the FSP matrix from two layers is a strict constraint which is hard to mimic well between student and teacher networks. At the same time, considering the structural property that pathology images exist pixel-wise similarity, the variation of cross-layer spatial similarity is employed to measure the correlation from different layers. Specifically, the channel-wise information in the teacher network can facilitate the performance of the student network [20]. Therefore, we mix channel-wise information into the spatial similarity which can further improve the performance of pathological gastric cancer segmentation. Also, inspired by using conditional Generative Adversarial Network in knowledge distillation [21], we utilize the adversarial learning strategy to once more improve the efficiency of our method. Furthermore, in the training procedure, we discover the paraphraser module [22] in an unsupervised way which can stable the process of knowledge transfer between teacher and student network.

Our main contributions can be summarized as follows:

• We propose CoCo DistillNet, a novel Cross-layer Correlation Distillation Network for pathological gastric cancer segmentation.

• The cross-layer correlation of channel-mixed spatial similarity is designed as a structural and robust knowledge which are transferred from teacher network to student network.

• Adversarial Distillation is utilized to further improve the performance through a training style, and the unsupervised Paraphraser Module is exploited to stabilize our training procedure.

• Extensive experiments are conducted on the Gastric Cancer Segmentation Dataset. Both qualitative and quantitative results validate the effectiveness of the proposed knowledge distillation method.

II. RELATED WORK

A. Pathology Image Segmentation

Pathology image segmentation dedicate to prompt diagnostic efficiency and accuracy by automatic labeling a corresponding category to each pixel in an image. Pathology image segmentation has witnessed impressive progress with the boost of deep neural networks. Ronneberger *et al.* [1] presented U-net that consists of an encoder to extract context and a decoder to enable precise segmentation. For accurate diagnosis of interstitial lung diseases, Anthimopoulos *et al.* [23] proposed a deep convolutional neural network which consists of dilated convolution. Bi*et al.* [24] designed a new stacked fully convolutional network with multi-channel learning and achieved great advances in colorectal cancer segmentation. Chan *et al.* [25] adopted a patch-based CNN in colon glands cancer segmentation, named HistoSegnet, which appears excellent performance and generalization. However, the inherent high-resolution pathological images prohibit deploying the networks aforementioned with millions of parameters in real scenes. Subsequently, it is urgent to propose a lightweight model.

B. Lightweight model in medical image analysis

Lightweight model aims to deploy networks which have high performance and low consumption in real scenes. In medical image analysis, there are various efforts to devise the lightweight model, which include designing lightweight networks, network pruning, network quantization, and knowledge distillation. A lightweight Spatial Attention U-net (SA U-net) was proposed by Guo *et al.* [26] for retinal vessel segmentation. Fernandes *et al.* [27] presented an automatic generating compact DNN used for medical image diagnostic with the help of network pruning. Zhang *et.al* [28] exploited network quantization technology to design a novel compact framework which obtain high effectiveness in several public 3D medical segmentation datasets. In the next, a few researchers utilized knowledge distillation to deal with the problem of efficiency in medical image analysis. Ho *et al.* [29] proposed a self-training knowledge distillation method to improve the accuracy in chest X-ray images classification. Li *et al.* [30] utilized online mutual knowledge distillation to improve the performance of CT image segmentation. Inspired by previous researches, we firstly introduce knowledge distillation in pathology image segmentation.

C. Knowledge Distillation

The target of knowledge distillation is to obtain a student network that has competitive or even a superior performance compared with the teacher network. Initially, Hinton *et al.* [10] regarded the 'soft target' of teacher network predictions as knowledge. Next, Ballas *et al.* [11] proposed the hints from the intermediate feature map of the teacher network, which can also improve the performance of the student network. Subsequently, the hints from the intermediate feature map were further extended. Tung *et al.* [12] raised a novel knowledge distillation method which is to preserve pairwise similarity in a batch. Yim *et al.* [19] treated the Flow of
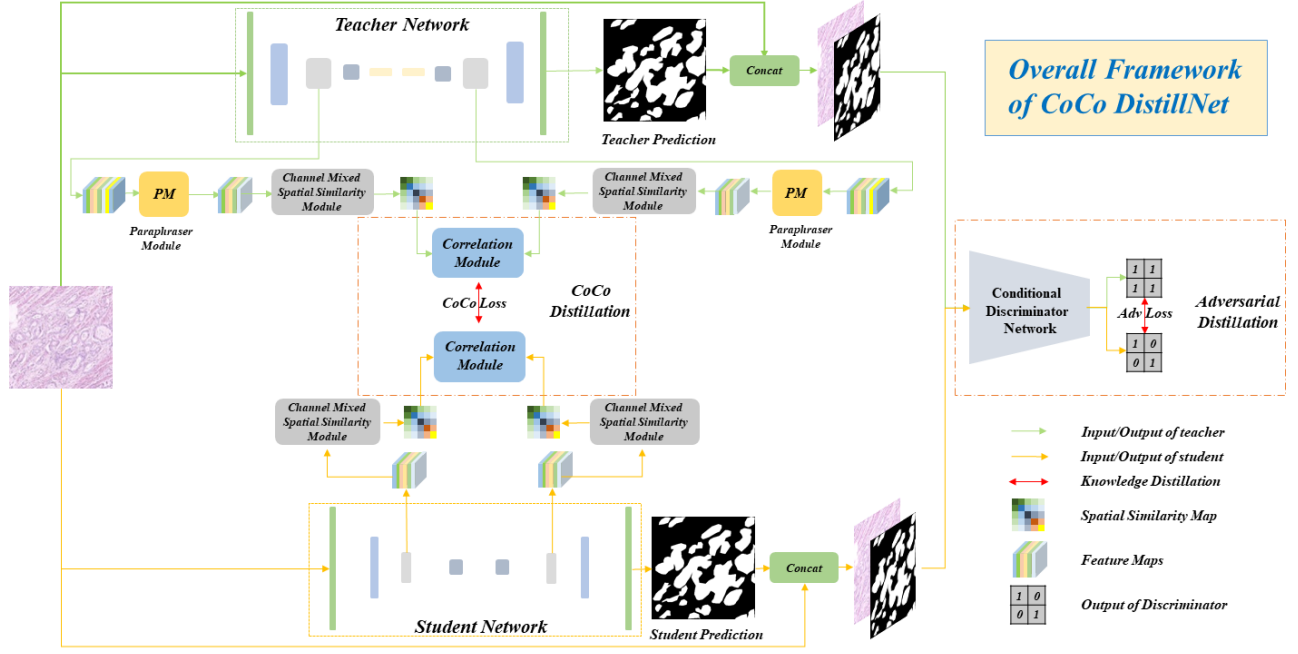
Fig. 1. The overall framework of CoCo DistillNet. Channel Mixed Spatial Similarity Module is committed to extract the pixel-wise similarity. Meanwhile, it could also assimilate channel-wise information. Correlation Module is utilized to compute the cross-layer correlation. The student network is supervised by CoCo Distillation and Adversarial Distillation. Paraphraser Module aims to paraphrase the knowledge from teacher network.

Solution Procedure (FSP matrix) as transferred knowledge and demonstrated the effectiveness in the CIFAR-100 dataset. He *et al.* [17] presented a method that minimizes the difference of pixel-wise affinity between teacher and student. Liu *et al.* [13] investigate the structured knowledge in semantic segmentation, and then exam the effectiveness of two structured distillation schemes, which are pairwise distillation and holistic distillation. Motivated by preceding studies, we consider the cross-layer correlation for knowledge transfer.

III. METHOD

Pathology image segmentation aims to predict a corresponding class to each pixel in a given pathology image X. Given a pre-trained teacher network T and a non-trained student network S, our CoCo DistillNet engages in improving the performance of S supervised by the knowledge from T. The outputs of S and T are noted as $Y_s = S(X)$ and $Y_t = T(X)$. we set $F_s^{l_s^j} \in R^{c \times h \times w} (j = 1, \cdots, i, \cdots, n+i, \cdots, 2n)$, where c is the number of channels, $h \times w$ is the spatial size, and n is the number of the selected all (l_s^i, l_s^{n+i}) pair layers, representing the feature maps of particular layer l_s^j produced by S. $F_t^{l_t^j} \in R^{c' \times h \times w} (j = 1, \cdots, i, \cdots, n+i, \cdots, 2n)$ denote the feature maps of particular layer l_t^j produced by T, where layer l_t^j is determined by the size of $F_s^{l_s^j}$. Note that c' is not necessarily to equal c due to the use of Parapharser Module. In this paper, we set n=1.

A. Overview

As illustrated in the Fig. 1, the pathology image X is input to the student network S and the teacher network T, simultaneously. Given the feature maps $F_t^{l_t^i}$ from layer l_t^i, the PM produces new feature maps $F_t^{l_t^i}{}'$ which can be easily comprehended by student network S. Furthermore, the Channel Mixed Spatial Similarity Module is exploited to acquire the channel-mixed spatial similarity hidden in $F_t^{l_t^i}{}'$ and $F_s^{l_s^i}$. Subsequently, the Correlation Module measures the correlation of the channel-mixed spatial similarity between $F_t^{l_t^i}{}'$ and $F_t^{l_t^{n+i}}{}'$; $F_s^{l_s^i}$ and $F_s^{l_s^{n+i}}$, respectively. For transferring the cross-layer correlation knowledge from S to T, we introduce the CoCo Loss to reduce the discrepancy of cross-layer correlation between S and T. Furthermore, the Adversarial Distillation is employed to enforce Y_s to mimic Y_t. At last, the student network S trained by our CoCo DistillNet can achieve competitive performance without any extra parameters and memory.

B. Cross-layer Correlation Distillation

1) Channel Mixed Spatial Similarity Module: Considering the pixel-wise similarity in pathology images, it is crucial to introduce the spatial similarity in pathology images segmentation. Furthermore, we propose the channel-mixed spatial similarity as each channel tends to encode different categories information in the feature maps. Concretely, as depicted in the Fig. 2, we use global max pooling to generate channel-wise information $A = [a_1, a_2, \cdots, a_c] \in R^{c \times 1 \times 1}$. Meanwhile, we reshape $F_{(\cdot)}^{l_{(\cdot)}^i}$ to $G_{(\cdot)}^{l_{(\cdot)}^i} \in R^{hw \times c}$ and exploit a softmax layer to normalize A. Next, we mix the channel-wise information A into $G_{(\cdot)}^{l_{(\cdot)}^i}$ to calculate $G_{(\cdot)}^{l_{(\cdot)}^i}{}'$, which refers to a $hw \times c$ matrix.
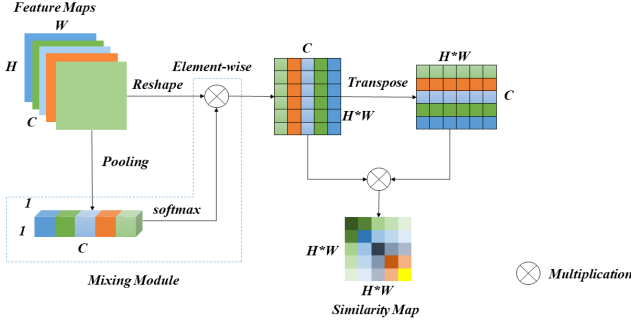
Fig. 2. The illustration of Channel Mixed Spatial Similarity Module. Mixing Module is proposed to assimilate channel-wise information into spatial similarity maps.

Fig. 3. The pipeline of Correlation Module. The cross-layer correlation is measured by similarity maps from different layers.

The computation could be formulated as:

$$g_{ij}' = g_{ij} \frac{\exp(a_j)}{\sum_{k=1}^{c} \exp(a_k)} \tag{1}$$

where $g_{ij}' \in G_{(\cdot)}^{l_{(\cdot)}^{i}}{}'$. Subsequently, we defined the channel-mixed spatial similarity map as $M_{(\cdot)}^{l_{(\cdot)}^{i}} \in R^{hw \times hw}$:

$$m_{ij} = \sum_{k=1}^{c} \frac{g_{ik}' g_{kj}''}{\sqrt{\sum_{z=1}^{c} g_{iz}'} \sqrt{\sum_{z=1}^{c} g_{zj}''}} \tag{2}$$

where $m_{ij} \in M_{(\cdot)}^{l_{(\cdot)}^{i}}$ refers to the similarity between i-th pixel and j-th pixel in the feature maps $F_{(\cdot)}^{l_{(\cdot)}^{i}}$, and g_{kj}'' is the element of $(G_{(\cdot)}^{l_{(\cdot)}^{i}}{}')^{T}$.

2) Correlation Module: We propose the Correlation Module to measure the correlation between low-level representations and high-level representations. As shown in Fig. 3, we utilize $M_{(\cdot)}^{l_{(\cdot)}^{i}}$ and $M_{(\cdot)}^{l_{(\cdot)}^{n+i}}$ to calculate the correlation $\phi(\cdot, \cdot)$ of representations between $l_{(\cdot)}^{i}$ and $l_{(\cdot)}^{n+i}$. The $\phi(\cdot, \cdot)$ can be formulated as:

$$\phi(F_{(\cdot)}^{l_{(\cdot)}^{i}}, F_{(\cdot)}^{l_{(\cdot)}^{n+i}}) = \frac{Q_{(\cdot)}^{l_{(\cdot)}^{i}} (Q_{(\cdot)}^{l_{(\cdot)}^{n+i}})^{T}}{\left\| Q_{(\cdot)}^{l_{(\cdot)}^{i}} \right\|_{2} \cdot \left\| Q_{(\cdot)}^{l_{(\cdot)}^{n+i}} \right\|_{2}} \tag{3}$$

where $Q_{(\cdot)}^{l_{(\cdot)}^{i}} \in R^{1 \times hwhw}$ is the reshaping of the $M_{(\cdot)}^{l_{(\cdot)}^{i}}$.

Next, we use Euclidean distance to minimize the discrepancy of the cross-layer correlation between teacher network and student network. We named this constraint as CoCo knowledge distillation loss function $\mathcal{L}_{coco}$ which could be presented as :

$$\mathcal{L}_{coco} = \frac{1}{n} \sum_{i=1}^{n} \left\| \phi(F_t^{l_t^{i}}, F_t^{l_t^{n+i}}) - \phi(F_s^{l_s^{i}}, F_s^{l_s^{n+i}}) \right\|_2^2 \tag{4}$$

3) Paraphraser Module: In the training procedure, the $\mathcal{L}_{coco}$ is difficult to converge stably and sufficiently. It indicates that
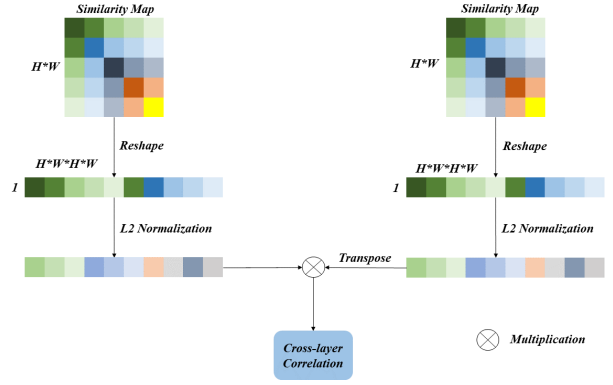
the knowledge from T is hardly transferred to S. Empirically, the knowledge embedded in T can not be comprehended well by S due to the feature distribution gap caused by mapping of T and S. Motivated by [22], we exploit the Paraphraser Module to embed the feature maps $F_t^{l_t^{i}}$ into the same feature space as $F_s^{l_s^{i}}$ without the loss of necessary knowledge from T. As illustrated in Fig. 4, the Paraphraser Module P is consist of two networks: encoder and decoder network which are stacked by several convolutional layers and transposed convolutional layers. Given the input $F_t^{l_t^{i}}$, the encoder network is used to compress $F_t^{l_t^{i}}$, while the decoder network aims to approximate $F_t^{l_t^{i}}$ as much as possible, so that the output of encoder network $P_{pm}(\cdot)$ can maintain the necessary knowledge. The encoder network $P_{pm}(\cdot)$ can be written as:

$$F_t^{l_t^{i}}{}' = P_{pm}(F_t^{l_t^{i}}) \tag{5}$$

where $F_t^{l_t^{i}}{}'$ is the output of $P_{pm}(\cdot)$. In the training mode, the reconstruction loss function $\mathcal{L}_{rec}$ is:

$$\mathcal{L}_{rec} = \left\| F_t^{l_t^{i}} - P(F_t^{l_t^{i}}) \right\|^2 \tag{6}$$

which is utilized to minimize the discrepancy between $F_t^{l_t^{i}}$ and $P(F_t^{l_t^{i}})$. For efficient minimizing the $\mathcal{L}_{coco}$, we utilize the paraphrasing mode to avoid the phenomenon aforementioned. Consequently, the $\mathcal{L}_{coco}$ can be modified as:

$$\mathcal{L}_{coco} = \frac{1}{n} \sum_{i=1}^{n} \left\| \phi(F_t^{l_t^{i}}{}', F_t^{l_t^{n+i}}{}') - \phi(F_s^{l_s^{i}}, F_s^{l_s^{n+i}}) \right\|_2^2 \tag{7}$$

C. Adversarial Distillation

In addition, we introduce the adversarial distillation to align the outputs between student network and teacher network. The student network S is deemed as a generator that aims to maximize the similarity between the prediction of student network S and output from teacher network T. As shown in Fig. 1, the discriminator D is trained to distinguish whether an input, which consists of the prediction Y_s/Y_t and the
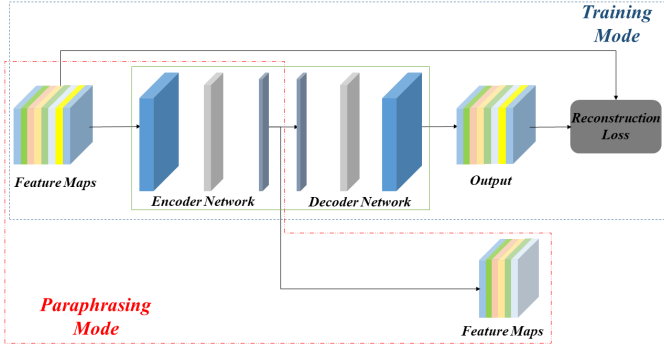
Fig. 4. The overall architecture of Paraphraser Module. The Training Mode is utilized to train the encoder and decoder network. In the Paraphrasing Mode, we only employ the encode network to paraphrase knowledge from teacher.

Fig. 5. The loss curves of $\mathcal{L}_{coco}$. Left: The loss curve of $\mathcal{L}_{coco}$ without the Paraphraser Module. Right: The loss curve of $\mathcal{L}_{coco}$ with the Paraphraser Module.

pathology image X, is from the teacher network or the student network. In the training procedure, the discriminator is supervised by the adversarial distillation function $\mathcal{L}_{ad}$, which can be formulated as:

$$\mathcal{L}_{ad} = E_{Y_t \sim \mathcal{P}_t(Y_t)}(D([Y_t, X])) - E_{Y_s \sim \mathcal{P}_s(Y_s)}(D([Y_s, X]) \tag{8}$$

where $[,]$ represents concatenation operation, $E(\cdot)$ indicates the expectation operation. $\mathcal{L}_{ad}$ is to minimize the wasserstein distance between Y_s and Y_t [18] [13]. For training the generator S, analogously with the training strategy of cGAN, the generator is trained by $\mathcal{L}_g$ which is written as:

$$\mathcal{L}_g = E_{Y_s \sim \mathcal{P}_s(Y_s)}(D([Y_s, X])) \tag{9}$$

D. Training Strategy

For optimizing our proposed overall CoCo DistillNet, the training process can be classified into three steps. To begin with, we leverage the $\mathcal{L}_{re}$ to train P, the parameters of T are frozen at the same time. Next, the $\mathcal{L}_{ad}$ is employed to train the discriminator network D which discriminates the input Y_s/Y_t as correct as possible. At last, the conventional cross-entropy loss $\mathcal{L}_{ce}$ is combined with $\mathcal{L}_{coco}$ and $\mathcal{L}_g$ to construct the fully objective $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{coco} - \beta\mathcal{L}_g \tag{10}$$

where λ and β are hyperparameters and we set λ and β to 1.0 and 0.1 respectively in this paper. Note that the optimizations of D and S are alternative implemented in the training procedure.

IV. EXPERIMENTS

A. Dataset

We employ the Gastric Cancer Segmentation Dataset [3] to validate the effectiveness of our proposed CoCo DistillNet. The Gastric Cancer Segmentation Dataset is sampled from clinical data. There are two semantic classes in this dataset including: normal area and cancerous area. The dataset is composed of 500 pathological images with the resolution of
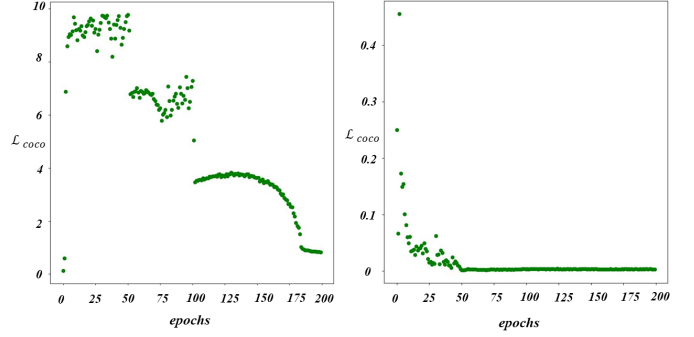
2048×2048, which are cropped from whole pathological slides of gastric regions [3]. We adjust the training and testing images and refer to the training strategy of [3]. In our experiments, we take patch-based method expressed as each 2048×2048 image is cropped into four 1024×1024 patches. Moreover, ordinary data augmentation methods are also employed in our training, such as randomly flip images horizontally and vertically.

B. Implementation Details

The FANet [31] which is a U-net based on self-attention (FANet) mechanism is employed as our teacher network T. Meanwhile, we adopt the U-net based on MobileNetV2 [33] (Mobile U-Net) as our student network S. Simultaneously, the lightweight networks ENet [34] and ERFNet [35] are chosen as our student network S to verify the generalization of our method.

For training S, T, and D, we use Adam as optimizer with weight decay 0.0002 and the 'ploy' learning rate policy $(1 - \frac{iter}{maxiter})^{power}$, where the $power$ is set to 0.9. For training P, we employ standard SGD as optimizer with the momentum 0.9, weight decay 0.0002 and the 'poly' learning rate policy. In the process of training S, T, and P, the learning rate is set to 0.03 initially and divided by 10 after every 50 epochs. The initial learning rate is set to 0.0002 for optimizing D. All of the networks is trained for 200 epochs on the Gastric Cancer Segmentation Dataset with the batch size of 8. We resize the input images size as 512×512 for the limited GPU memory. All experiments are implemented on the Pytorch platform with 2 NVIDIA 2080TI GPUs.

C. Evaluation Metrics

In all experiments, we utilize the commonly used pixel accuracy (ACC) and mean Intersection-over-Union (mIoU) to evaluate the performance of our networks. The ACC is calculated as the ratio of the correctly predicted pixels and total pixels. The mIoU is computed as the mean of the ratio of

interval and union between the ground truth and the prediction of all classes. The ACC and mIoU can be formulated as:

$$ACC = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \quad (11)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \quad (12)$$

where $k+1$ refers to the number of classes, p_{ii} represents the number of pixels which are belong to i-th class and predicted for i-th class, and p_{ij} represents the the number of pixels which are belong to i-th class and predicted for j-th class. We also utilize the number of parameters (Params) and the sum of point operations (FLOPs) to measure model complexity.

TABLE I
ABLATION STUDY IN THE GASTRIC CANCER SEGMENTATION DATASET SS: SPATIAL SIMILARITY. CMSS: CHANNEL MIXED SPATIAL SIMILARITY MODULE. CoCoD: CoCo DISTILLATION. AdvD: ADVERSARIAL DISTILLATION. CS AND SS REFERS TO ONLY TRANSFER THE SPATIAL SIMILARITY WITH OR WITHOUT THE CHANNEL-WISE INFORMATION FROM T TO S.

Method				ACC	mIoU
T: FANet				0.9029	0.8230
S: Mobile U-net				0.8555	0.7475
SS	CMSS	CoCoD	AdvD		
✓				0.8615	0.7567
	✓			0.8656	0.7630
		✓		0.8739	0.7761
		✓	✓	**0.8780**	**0.7824**

D. Ablation Study

1) Module Ablation: We conduct extensive experiments to verify the effectiveness of each component of our proposed CoCo DistillNet. We conduct four set of experiments for module ablation: Spatial Similarity (SS), Channel Mixed Spatial Similartiy (CMSS), CoCo Distillation (CoCoD) and Adversarial Distillation (AdvD). SS and CMSS are utilized to verify the effectiveness of our proposed Channel Mixed Spatial Similarity module. The distinction between SS and CMSS is that the channel-wise information is mixed or not mixed into the spatial similarity. CoCoD is employed to exam our proposed cross-layer correlation of channel-mixed spatial similarity. Furthermore, we apply CoCoD and AdvD to demonstrate the effectiveness of our complete method. These results are shown in Table. I. From Table. I, we observe that the spatial similarity improve the S without distillation by 0.60% (ACC) / 0.92% (mIoU) and the Channel Mixed Spatial Similarity module can further increase the performance of the S without distillation to 1.01% (ACC) / 1.55% (mIoU). It demonstrates that our proposed Channel Mixed Spatial Similarity module can extract spatial similarity and channel information effectively. With the use of our CoCo Distillation, the improvements of the S are 1.84% (ACC) / 2.86% (mIoU), indicating the knowledge defined as the cross-layer correlation of channel-mixed spatial similarity can significantly prompt

the performance of the S. At last, the total CoCo DistillNet, i.e. with the use of CoCo Distillation and Adversarial Distillation, can prompt the ACC and mIoU of the S to reach 87.80% and 78.24% respectively. Fig. 6 shows the qualitative results of module ablation. We also demonstrate the effectiveness of the Paraphraser Module in Fig. 5. We observe that the $\mathcal{L}_{coco}$ curve appears irregular vibration and significantly reduction after almost every 50 epochs without Paraphraser Module as shown in Fig. 5 (left). As illustrated in Fig. 5 (right), Paraphraser Module could effectively enhance training stabilization and converge rapidly.

TABLE II
COMPARISON WITH DIFFERENT KNOWLEDGE DISTILLATION METHODS.

Method	ACC	mIoU
S:Mobile U-Net	0.8555	0.7475
S+KD [10]	0.8620	0.7574
S+AT [32]	0.8679	0.7665
S+FSP [19]	0.8636	0.7600
S+IFVD [16]	0.8682	0.7671
S+CoCo(ours)	**0.8780**	**0.7824**

2) Comparisons with other knowledge distillation methods: As illustrated in the Table. II, our proposed CoCo DistillNet achieves the state-of-the-art performance compared with several popular knowledge distillation methods. Concretely, our method outperforms KD by 1.60% (ACC) / 2.50% (mIoU), and AT by 1.01% (ACC) / 1.59% (mIoU), in which KD and AT are both designed for classification task. It indicates the superiority of our method in pathology image segmentation. Compared with FSP, our methods outperforms the performance of the FSP by 1.44% (ACC) / 2.24% (mIouU). This is because our method constructs more appropriate cross-layer correlation than FSP. Our method also surpasses IFVD by 0.98% (ACC) / 1.53% (mIoU). It demonstrates that our method utilizes the structural property of pathology images and the correlation between shallow and depth layers can effectively instruct the training of S, better than other methods devised for segmentation task.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT STUDENT NETWORKS. MUNET REPRESENTS MOBILE U-NET. CoCo REFERS TO THAT THE STUDENT NETWORK TRANINED BY OUT CoCo DISTILLNET.

Type	Network	ACC	mIoU	FLOPs	Params
T	FANet	0.9029	0.8230	171.556G	38.250M
S	ERFNet	0.8695	0.7691	3.681G	2.063M
	ERFNet+CoCo	**0.8779**	**0.7824**		
	MUNet	0.8555	0.7475	1.492G	4.640M
	MUNet+CoCo	**0.8780**	**0.7824**		
	ENet	0.8766	0.7802	0.516G	0.349M
	ENet+CoCo	**0.8840**	**0.7919**		

E. Results

The effectiveness of our method is examined with several compact networks: Mobile U-Net, ERFNet, and ENet. The Table. III represents the performances (ACC and mIoU),
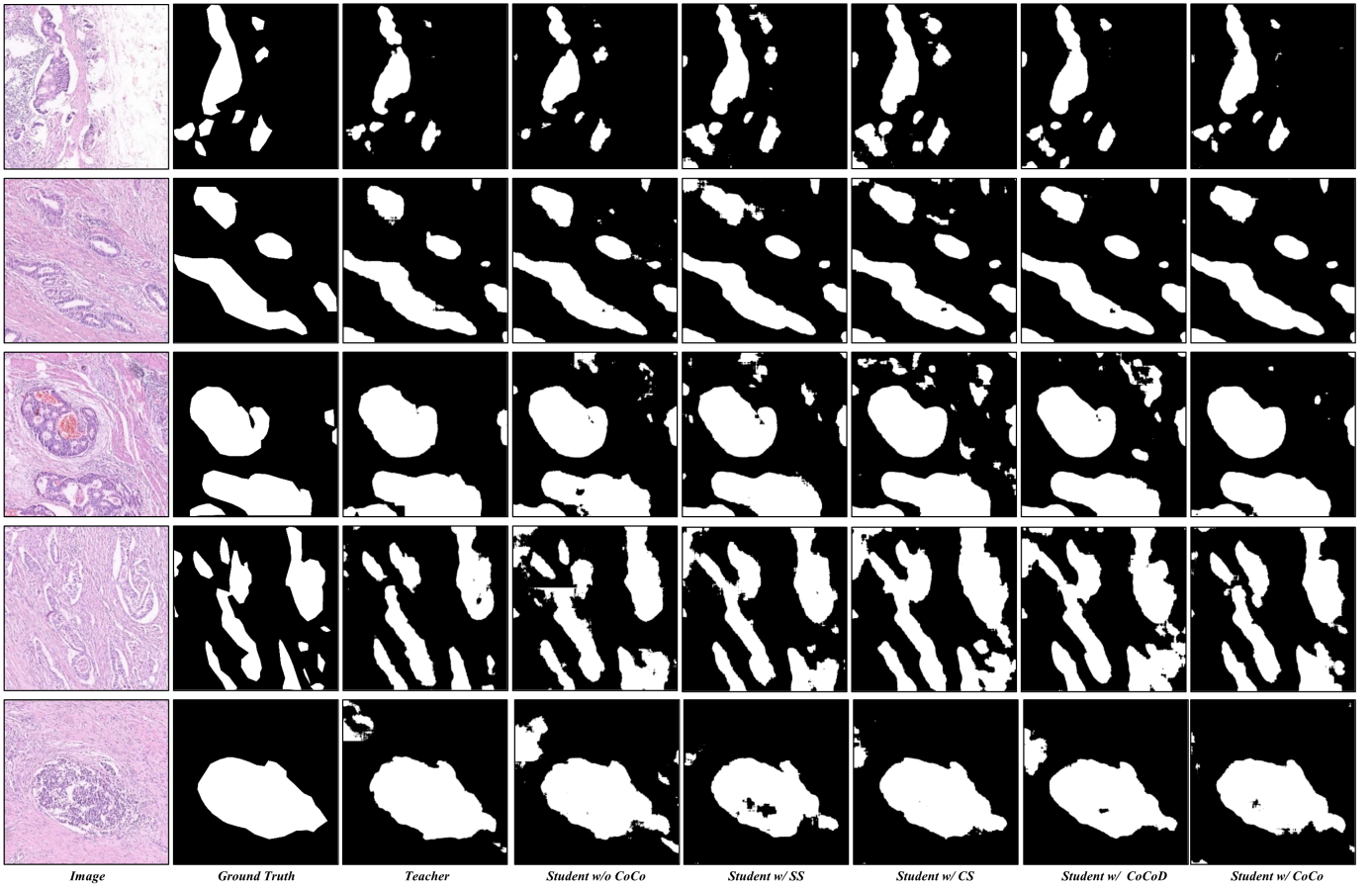
Fig. 6. Qualitative results of our proposed method.Teacher refers to FANet. Student w/o CoCo represents the Mobile U-Net trained from scratch. SS: Spatial Similarity. CMSS: Channel Mixed Spatial. Similarity CoCoD: CoCo Distillation. CoCo: CoCo DistillNet (CoCo Distillation+Adversarial Distillation).

complexity (FLOPs), and size (Params) of these networks. The Mobile U-Net, ERFNet, and ENet are all lightweight networks in which the Mobile U-Net has the similarity architecture with the FANet. In contrast, the ENet and ERFNet have different architecture with the FANet. As shown in Table.III, there are huge gaps between the performance of these lightweight networks and the cumbersome network (FANet). However, the gains of Mobile U-Net can achieve a large margin 2.25% (ACC) / 3.49% (mIoU) when employing our method. For ERFNet, we can see that the ACC is improved by 0.84%, and the mIoU is improved by 1.33%. At the same time, our method can boost the ENet to achieve 88.40% (ACC) / 79.19% (mIoU), which is a approximate performance compared with the FANet. With our proposed method, The gap between the teacher network and these compact network is alleviated significantly. It testifies that our method can improve the performance of compact networks whether the S and the T have similar architectures or not.

Furthermore, we also conduct qualitative analysis. As shown in Fig. 6, we select several pathology images with labels and show their predictions from different settings. In these experiments, the Mobile U-Net is applied as the student network S. We can see that the student network with our method can

distinguish the cancerous regions more accurately than the student network without distillation. Moreover, our method can reduce the error of being mispredicted as cancerous regions even compared with the teacher network T.

V. Conclusion

In this paper, we present a novel knowledge distillation method tailored for pathology image segmentation. We propose the channel-mixed spatial similarity based on previous researchers. Furthermore, we define the Cross-layer Correlation of channel-mixed spatial similarity as a structural and robust knowledge. We use the knowledge transferred from the cumbersome network T to improve the performance of the compact network S. The Adversarial Distillation is employed to further improve the performance of our method. Furthermore, we also use the Praphraser Module to stabilize our training. Our method achieves the state-of-the-art performance compared with various knowledge distillation methods. In addition, we also utilize different compact networks to verify the generalization of our method. In the future, we would like to extend our method in other multi-class and multi-model medical image segmentation tasks. We also expect

to generalize the channel-mixed spatial similarity to further reduce computation cost.

ACKNOWLEDGMENT

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, "Adaptive weighting multi-field-of-view cnn for semantic segmentation in pathology," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 597–12 606.

[3] M. Sun, G. Zhang, H. Dang, X. Qi, X. Zhou, and Q. Chang, "Accurate gastric cancer segmentation in digital pathology images using deformable convolution and multi-scale embedding networks," *IEEE access*, vol. 7, pp. 75 530–75 541, 2019.

[4] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[5] J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X.-s. Hua, "Quantization networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7308–7316.

[6] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, 1990, pp. 598–605.

[7] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," *arXiv preprint arXiv:1506.02626*, 2015.

[8] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4820–4828.

[9] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[11] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[12] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[13] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[14] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[15] Z. Chen, X. Zheng, H. Shen, Z. Zeng, Y. Zhou, and R. Zhao, "Improving knowledge distillation via category structure," in *European Conference on Computer Vision*. Springer, 2020, pp. 205–219.

[16] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, and C. C. Loy, "Inter-region affinity distillation for road marking segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 486–12 495.

[17] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 578–587.

[18] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–362.

[19] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[20] C. Shu, Y. Liu, J. Gao, L. Xu, and C. Shen, "Channel-wise distillation for semantic segmentation," *arXiv preprint arXiv:2011.13256*, 2020.

[21] Z. Xu, Y.-C. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," *arXiv preprint arXiv:1709.00513*, 2017.

[22] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," *arXiv preprint arXiv:1802.04977*, 2018.

[23] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, and S. Mougiakakou, "Semantic segmentation of pathological lung tissue with dilated fully convolutional networks," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 714–722, 2018.

[24] L. Bi, J. Kim, A. Kumar, M. Fulham, and D. Feng, "Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation," *The Visual Computer*, vol. 33, no. 6, pp. 1061–1071, 2017.

[25] L. Chan, M. S. Hosseini, C. Rowsell, K. N. Plataniotis, and S. Damaskinos, "Histosegnet: Semantic segmentation of histological tissue type in whole slide images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 662–10 671.

[26] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, "Sa-unet: Spatial attention u-net for retinal vessel segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1236–1242.

[27] F. E. Fernandes and G. G. Yen, "Automatic searching and pruning of deep neural networks for medical imaging diagnostic," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[28] R. Zhang and A. C. Chung, "Medq: lossless ultra-low-bit neural network quantization for medical image segmentation," *Medical Image Analysis*, p. 102200, 2021.

[29] T. K. K. Ho and J. Gwak, "Utilizing knowledge distillation in deep learning for classification of chest x-ray abnormalities," *IEEE Access*, vol. 8, pp. 160 749–160 761, 2020.

[30] K. Li, K. Yu, S. Wang, and P.-A. Heng, "Towards cross-modality medical image segmentation with online mutual knowledge distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 775–783.

[31] K. Li, X. Qi, Y. Luo, Z. Yao, X. Zhou, and M. Sun, "Accurate retinal vessel segmentation in color fundus images via fully attention-based networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2071–2081, 2020.

[32] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[34] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[35] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.