Graph Flow: Cross-Layer Graph Flow Distillation for Dual Efficient Medical Image Segmentation

Wenxuan Zou\textsuperscript{a,b}, Xingqun Qi\textsuperscript{c,d}, Wanting Zhou, Muyi Sun\textsuperscript{e,f}, Zhenan Sun\textsuperscript{g}, Senior Member, IEEE, and Caifeng Shan\textsuperscript{a}, Senior Member, IEEE

Abstract—With the development of deep convolutional neural networks, medical image segmentation has achieved a series of breakthroughs in recent years. However, high-performance convolutional neural networks always mean numerous parameters and high computation costs, which will hinder the applications in resource-limited medical scenarios. The scarcity of large-scale annotated medical image datasets further impedes the application of high-performance networks. To tackle these problems, we propose Graph Flow, a comprehensive knowledge distillation framework, for both network-efficiency and annotation-efficiency medical image segmentation. Specifically, the Graph Flow Distillation transfers the essence of cross-layer variations from a well-trained cumbersome teacher network to a non-trained compact student network. In addition, an unsupervised Paraphraser Module is integrated to purify the knowledge of the teacher, which is also beneficial for the training stabilization. Furthermore, we build a unified distillation framework by integrating the adversarial distillation and the vanilla logits distillation, which can further refine the final predictions of the compact network. With different teacher networks (traditional convolutional architecture or prevalent transformer architecture) and student networks, we conduct extensive experiments on four medical image datasets with different modalities (Gastric Cancer, Synapse, BUSI, and CVC-ClinicDB). We demonstrate the prominent ability of our method on these datasets, which achieves competitive performances. Moreover, we demonstrate the effectiveness of our Graph Flow through a novel semi-supervised paradigm for dual efficient medical image segmentation. Our code will be available at Graph Flow.

Index Terms—Knowledge distillation, efficient medical image segmentation, semi-supervised learning, graph flow.

1. INTRODUCTION

MEDICAL image segmentation (MIS), a challenging computer vision task, aims to automatically assign pixel-wise labels for the specific regions in medical images, such as organs, cells, lesion areas, as precisely as possible. In the field of computer-aided diagnosis and therapy, MIS plays a crucial role, which can efficiently alleviate the burden for clinicians to identify pathological or anatomical regions. In the past decade, the employment of Convolutional Neural Networks (CNNs) has become a milestone for medical image segmentation. Various elaborate CNNs \cite{1,2,3,4,5} exhibit superior performances on medical images of different modalities. Such tailored CNNs always consist of sophisticated network architectures with millions of parameters and vast computational operations. However, these networks are degraded or even incompetent if employed in the scenes with limited computational resources such as in the rural clinics of developing countries, or in the real-time diagnosis scenario of large hospitals with tens of thousands of patients daily. These situations are further aggravated when processing the medical images with huge resolution and deep volumes (e.g. pathological Whole-Slide-Image \cite{6,7} and Computed Tomography (CT) \cite{8,9}), or operating at bedsides on mobile devices (e.g. ultrasound image \cite{10} or colonoscopy image \cite{11}). In addition, the annotation of medical images is expensive due to the requirements of experienced experts and abundant time. Thus, large-scale annotated medical image datasets are always scarce. The above characteristics of the CNNs and medical images result in that the high performance networks with less computation and storage costs are unavailable in resource-limited and label-limited medical application scenes. Therefore, the trade-off between the performance and cost of CNNs needs to be considered. At the same time, the massive and available unlabeled medical data should also be taken into account as much as possible, due to the lack of annotated medical images in real applications.

To tackle the resource-limited dilemma, some researchers engage in devising lightweight networks in the medical imaging community, since lightweight networks consume less computational resource and inference time than cumbersome networks. Zhou et al. \cite{12} devise a double attention network with a pre-trained MobileNetV2 encoder for real-time...
Recenty, knowledge distillation, a straightforward and effective model compression technique, has drawn more and more attention in computer vision. Knowledge distillation is to distill and transfer the important knowledge hidden in a pre-trained cumbersome network (teacher network) to a non-trained compact network (student network) [15]. Meanwhile, knowledge distillation can promote the existing compact networks to extract more informative features than themselves trained from scratch, without redesigning delicate compact architectures. In the field of computer vision, the initial idea of knowledge distillation is to mimic the softened logit distributions between the teacher network and the student network [16]. Subsequently, researchers introduce knowledge distillation into the intermediate features of neural networks by direct approximating the features of student networks [17]. Nevertheless, it is difficult for the student network to mimic the intermediate representations of the teacher network with different network architecture. To this end, more studies based on relation knowledge (e.g. class-wise and layer-wise relation) are proposed [18], [19], [20] recently. As for class-wise relation distillation [19], [20], these methods model the pixel-level correlations between classes but lack holistic class-wise structural information, which might damage the exact segmentation of crucial areas. On the other hand, the layer-wise relation distillation [18] would transfer various redundant knowledge.

Though the knowledge distillation has developed deeply in computer vision, only a few researchers try to explore knowledge distillation in the medical imaging community. Wen et al. [21] propose a boundary-guided distillation method for organ segmentation. Qin et al. [22] design a region affinity distillation method for efficient liver tumor segmentation. Tran et al. [23] propose a new Light-weight Deformable Registration network that introduces a new adversarial learning with distilling knowledge. Since their methods are primary explorations for utilizing knowledge distillation in medical imaging analysis, they do not notice the variations of pathological or organic representations from different layers.

In this paper, considering the properties of medical datasets, segmentation network architectures, and the mentioned drawbacks of knowledge distillation, we propose a novel knowledge distillation framework, Graph Flow, to promote the trade-off between the performance and efficiency in efficient MIS. It worth emphasizing that the mentioned ‘efficiency’ involves two meanings: network efficiency and annotation efficiency. In other words, our method could not only compress the complexity of high-performance networks to satisfy the requirement of resource-limited medical scenarios, but also exploit the massive unlabeled medical images to alleviate the lack of annotations.

We visualize the motivation of our method on the Synapse Multi-organ Segmentation Dataset, as shown in Fig. 1. The feature flow of cross-layers exhibits the procedures of solution questions of CNNs, which is more critical than teaching students to remember answers. However, the flow of cross-layers contains redundant or useless knowledge, which may confuse the students to comprehend such procedures well. As illustrated in Fig. 1, the essence of cross-layer flow is the variation of structural semantic-aware knowledge, while the semantic-aware knowledge can be represented well based on an effective graph. Specifically, we firstly filter out the patch-wise salience regions based on the maximum activation area. Next, we suppress the activations of regions except for these salience regions along the axis of feature channels. After that, we propose our salience graph to encode structural semantic-aware knowledge without any other useless knowledge beyond these channel-wise and patch-wise salience regions. At last, we construct our variation graph to measure the flow of salience graphs between different layers. Therefore, our method, called Graph Flow Distillation, distills the variation graph from teacher networks to student networks. In addition, we utilize a Paraphraser Module [24] to refine the representations of teacher network. Moreover, we integrate the adversarial training strategy [25] (Adversarial Distillation) and the vanilla knowledge distillation [16] (Logits Distillation) into our Graph Flow to establish a full-stage knowledge distillation framework, which are both beneficial to the final performance.
On account of the intrinsic characteristics of knowledge distillation, our method could also transfer knowledge from a cumbersome teacher (pre-trained with the limited labeled data) without the supervision of annotation. Thus, our Graph Flow could be further utilized to conquer the scarcelness of annotations for efficient MIS in a semi-supervised manner.

Our main contributions are as follows:

- We propose Graph Flow Distillation, a novel knowledge distillation method, to transfer the flow of cross-layer salience graphs from a teacher network to a student network.
- We propose a unified distillation framework, dubbed Graph Flow, to incorporate Graph Flow Distillation, Adversarial Distillation, and Logits Distillation, and further provide a new paradigm for semi-supervised efficient medical image segmentation based on our Graph Flow with a tweaking on the objective functions.
- We conduct extensive experiments on the different-modality and multi-category medical image segmentation datasets. Extensive quantitative and qualitative analyses demonstrate the state-of-the-art ability of our Graph Flow.

This work has several noteworthy extensions compared with our preliminary work [26]: 1) We distill the flow of cross-layer salience graphs by variation graph instead of the variations categories. 4) We verify our method on more medical image segmentation datasets with different modalities and multiple categories. 4) We provide a new paradigm for semi-supervised efficient medical image segmentation.

II. RELATED WORK

A. Medical Image Segmentation

Medical image segmentation engages in automatic labeling organic or pathological structures in medical images. Recently, medical image segmentation has achieved impressive advances with the development of deep learning. U-Net [1], consisting of a symmetric architecture and skip connections, has served as a backbone network in medical image segmentation. Zhou et al. [3] design UNet++ with a nested architecture and dense-skip connections for improving the details of medical image segmentation. Milletari et al. [2] present V-net which utilizes 3D convolution and residual connections for volumetric medical data. Isensee et al. [4] propose a novel pipeline from self-configuring pre- and post-processing of dataset to U-shape network architecture. Though the above networks achieve the state-of-the-art performance in different modalities of medical images, they are not applicable in harsh medical scenes due to the incredible high storage, computational, and annotation cost. Therefore, some researchers intend to devise lightweight network architectures. Zhang et al. [28] devise a lightweight convolutional neural network with 2D and 3D convolution jointly. Guo et al. [29] propose a lightweight Spatial Attention U-net for retinal vessel segmentation. Nevertheless, such compact networks result in performance degradation with large-scale networks. It is urgent to propose an efficient method which can circumvent the trade-off between the performance and efficiency in medical image segmentation.

B. Knowledge Distillation

The knowledge distillation can be briefly defined as the training of a minor student network with the beneficial knowledge transferred from a cumbersome teacher network. Hinton et al. [16] firstly propose the notation of knowledge distillation and demonstrate that the soft targets of the teacher network can prompt the performance of the student network. Next, Romero et al. [17] propose FitNet using intermediate hidden layers of a teacher with a guided layer. However, such straightforward feature-based distillation methods are not sufficient for generalization. In the field of image segmentation, most knowledge distillation studies focus on relation-based knowledge. Liu et al. [20] utilize pair-wise distillation and holistic distillation to guide the learning of the student. Shu et al. [30] mimic the Kullback–Leibler divergence of the channel-wise probability maps between the teacher and the student. Hou et al. [31] design an inter-region graph based on the similarity for road marking segmentation. Meanwhile, other previous works also explore the graph-based methods [32], [33], [34]. Liu et al. [32] design a relationship graph (IRG) between the images (instances) in the same batch to measure the intra-class variations and the inter-class differences. Zhou et al. [33] construct a unified graph-based embedding, called attributed graph, by aggregating individual knowledge from relational neighborhood images. Lee et al. [34] propose Multi-Head Graph Distillation (MHGD) for transferring the intra-data relation knowledge between samples in the mini-batch. However, these methods are all designed based on instance-level features for image classification; compared to the semantic-aware features in our method, the instance-level features are redundant and can not focus on specific semantic categories. Besides, only a few researchers introduce knowledge distillation for efficient medical image segmentation. Wen et al. [21] propose a boundary-guided knowledge distillation method which assists the student network to align organ boundary features in teacher network. However, the effectiveness of this method depends on the accuracy of boundary extraction. Qin et al. [22] design a region affinity distillation method, and integrate the importance map for efficient liver tumor segmentation, while this method perform unsatisfactory due to the mismatching of auxiliary masks and feature maps. Moreover, these methods mostly concentrate on distilling the knowledge from certain layers, while do not notice the variations of pathological or organic semantic-aware representations from different layers. Inspired by the mentioned knowledge distillation methods and the intrinsic property of medical images, we propose Graph Flow for dual efficient medical image segmentation.

C. Semi-Supervised Medical Image Segmentation

The achieved remarkable successes of deep learning in medical image segmentation are widely dependent on massive high-quality annotated datasets. However, such datasets are usually unavailable since medical datasets are expensive...
Fig. 2. The overall framework of our Graph Flow. As a comprehensive knowledge distillation framework, it integrates three different strategies: Graph Flow Distillation, Adversarial Distillation, and Logits Distillation. In Graph Flow, Paraphraser Module is committed to stable and refine the knowledge transfer. In this framework, the network with yellow background represents the teacher network, the network with green background represents the student network, and the light pink background area represents the distillation operations.

to annotate [35]. Semi-supervised learning can relieve the scarcity of dataset annotation by exploring unlabeled data. Yu et al. [36] propose a semi-supervised medical segmentation framework based on uncertainty-awareness in which an uncertainty consistent loss can prompt the accuracy of segmentation. Yang et al. [37] design a novel hybrid loss function for semi-supervised medical instrument segmentation, consisting of a uncertainty and contextual constraints. Nevertheless, current semi-supervised medical image segmentation methods do not take account of the limitation of computational resources and memory storage in real scenes, while our Graph Flow can solve the lack of annotated datasets and the limitation of computational resources and memory storage simultaneously. Therefore, our semi-supervised strategy is embedded in our distillation method for dual efficient medical image segmentation.

III. METHOD

The purpose of our Graph Flow is to improve the performance of the non-trained compact target student network $S$ under the extra supervision of the knowledge transferred from the cumbersome teacher network $T$ in the medical image segmentation task. The pipeline of our Graph Flow is illustrated in Fig. 2, which consists of three different distillation methods including Graph Flow Distillation, Adversarial Distillation, and Logits Distillation. Our Graph Flow Distillation takes into account cross-layer knowledge flow, as shown in Fig. 1, which is neglected by previous knowledge distillation methods in medical image segmentation. Meanwhile, we also introduce Adversarial Distillation and Logits Distillation that constrain knowledge consistency between the outputs of $T$ and $S$ combined with distilling knowledge from intermediate layers to fulfill the full-stage distillation. Moreover, our method can simultaneously solve resource-limited and label-limited dilemmas of medical image segmentation in clinical diagnosis. Therefore, our Graph Flow is a comprehensive framework for dual efficient medical image segmentation.

To begin with, given a medical image $X$ as input, the teacher network $T$ which is pre-trained by ourselves produces features maps $F_i$ from $i_{th}$ layer, and outputs $Y_t$. Next, we employ the mentioned distillation methods to transfer different knowledge hidden in intermediate layers and outputs of $T$ for training the student network $S$ with the same input. Concretely, our Graph Flow Distillation constructs and matches variation graphs between $S$ and $T$ for distilling the cross-layer knowledge flow. In addition, Adversarial Distillation utilizes a discriminator network $D$ to align high-order knowledge between $Y_s$ and $Y_t$. At the same time, Logits Distillation is employed to mimic the logits distributions between $T$ and $S$. At last, the student
network $S$ trained by our Graph Flow can achieve competitive performance compared with the teacher network $T$ without any extra memory and parameters.

### A. Graph Flow Distillation

As illustrated in Fig. 1, the attention of medical image segmentation network greatly changes between different layers. Therefore, such knowledge hidden in the flow of cross-layers of teacher network can teach student network to extract more critical features for accurate organic or pathological segmentation. However, previous layer-wise relation distillation [18] is inappropriate for medical image segmentation because it simply calculates the relation between different layers so that does not focus on the variations of semantic-aware knowledge. Inspired by this phenomenon, we propose our salience graph, which can focus more on structural semantic-aware knowledge (class-wise spatial information and inter-class correlation) of an intermediate layer. Our salience graph can distinguish ambiguous semantic categories well, which is beneficial for capturing class-wise spatial information and inter-class correlation. Furthermore, we devise a variation graph, which can simultaneously represent the flow of class-wise spatial information and inter-class correlation embedded by salience graph from cross-layers.

### 1) Salience Graph

For feature maps $F(i)$ with the size of $C \times H \times W$ ($C$ is the number of channels, $H \times W$ is the spatial size), we propose salience graph denoted as $SG(i)$ ($\cdot$ refers to $s$ or $t$) to represent structural semantic-aware knowledge hidden in channel-wise salience features. The construction of $SG(i)$ is depicted in Fig. 3, which could be formulated as:

$$SG(i) = (\Gamma(i), \Theta(i)),$$

$$\Gamma(i) = MF(i),$$

$$\Theta(i)(c, k) = \Phi(\Gamma(i)(c), \Gamma(i)(k))(c, k = 1, 2, \cdots, C),$$

where $\Gamma(i)$ and $\Theta(i)$ are the vertex and edge set of $SG(i)$ respectively. $\Phi(\cdot, \cdot)$ is a function for measuring the correlation between different channel-wise salience regions. Each vertex in the $SG(i)$ represents one of the channels in the $F(i)$. As shown in Equation 2, we utilize $M = [M_1, M_2, \cdots, M_c, \cdots, M_C]$ to filter the redundant activations except for the patch $P = [P_1, P_2, \cdots, P_c, \cdots, P_C]$ centered on the pixel with the largest activation in every channel. The $M$ is a binary mask with the same size as $F(i)$. The value of each pixel of $M_i$ is computed by $M_i(p) = \mathbb{1}(p \in P_i)(p = 1, 2, \cdots, H \times W)$, where $\mathbb{1}()$ is an indicator function, and $p$ is a pixel of $c_{ih}$ channel of $F(i)$. The different sizes of $P$ indicate that salience graph $SG$ has different sizes of respective field about class-wise knowledge. After that, each vertex can encode a specific category with holistic spatial information. Each edge of $SG(i)$, $\Theta(i)(c, k)$, are computed by a score function $\Phi(\cdot, \cdot)$, which can encode inter-class correlation and does not destroy holistic structure of the category. We simply define the score function $\Phi(\cdot, \cdot)$ as Euclidean distance in this paper.

### 2) Variation Graph

Assume that we have a pair salience graphs, $SG(i)$ and $SG(i')$. We design a novel variation graph to measure the flow of salience graphs from $i_{th}$ layer to $i'_{th}$ layer, in which the $i'_{th}$ layer is a corresponding layer of $i_{th}$ layer with same size $H \times W \times C$. The variation graph, termed $VG(i)$, can be presented as:

$$VG(i) = (V(i), E(i)),$$

$$V(i)(c) = \|\Gamma(i)(c) - \Gamma(i')(c)\|_2^2 (c = 1, 2, \cdots, C),$$

$$E(i)(c, k) = \|\Theta(i)(c, k) - \Theta(i')(c, k)\|_2^2 (c, k = 1, 2, \cdots, C)$$

where $V(i)$ and $E(i)$ are the vertex and edge set of $VG(i)$ respectively. Each vertex of $VG(i)$, $V(i)(c)$, represents the changes of the same vertex of different salience graphs. In other words, $V(i)(c)$ can measure the spatial variations of a specific category from a shallow layer to a deep layer. The edge of $VG(i)$, $(i, E(i)(c, k))$ is employed to evaluate the variations of inter-class correlation from a shallow layer to a deep layer.

For transferring the knowledge of variations graphs, we devise variation graph loss function $L_{vg}$ to mimic the discrepancy of variation graphs between $T$ and $S$. The objective function $L_{vg}$ is defined as follows:

$$L_{vg} = \lambda_1 L_{vertex} + \lambda_2 L_{edge}$$

$$\lambda_1 \frac{1}{2|L|C} \sum_{c=1}^{C} \sum_{i \in L} \|V_i(c) - V_s(c)\|_2^2$$

$$\lambda_2 \frac{1}{2|L|C^2} \sum_{i \in L} \sum_{c=1}^{C} \sum_{k=1}^{C} \|E_i(c, k) - E_s(c, k)\|_2^2$$

(7)
where \( L \) notes the set of \((t_i, t'_i)\) pair layers used to construct salience graphs. The \(|L|\) is the size of \( L \), which indicates the number of pair layers. We set \(|L| = 1\) in this paper.

**B. Paraphraser Module**

It is worth noting that in our Graph Flow Distillation, the \( F_{ti} \) and \( F_{si} \) always have different feature channels due to the different network architectures of \( T \) and \( S \). Thus, the knowledge embedded in \( T \) can not be comprehended well by \( S \), due to the different latent spaces of \( F_{ti} \) and \( F_{si} \).

Motivated by [24], we utilize a Paraphraser Module \( P_m \) to ease the latent space gap between \( F_{ti} \) and \( F_{si} \) for refining the knowledge from \( T \) and stabilizing the training of \( S \). The overall architecture of our Paraphraser Module \( P_m \) is illustrated in Fig. 4. Specifically, our Paraphraser Module \( P_m \) consists of a encoder network \( P_{en} \) and a decoder network \( P_{de} \). The encoder network \( P_{en} \) is stacked by three \( 3 \times 3 \) convolutional operations with stride = 1 and padding = 1. The decoder network \( P_{de} \) is stacked by three corresponding transposed convolutional layers. The leaky-ReLU is used after the convolutions and transposed convolutions. Given an input \( F_{ti} \), the encoder network \( P_{en} \) aims to compress \( F_{ti} \) into the same latent space of \( F_{si} \), while the decoder network \( P_{de} \) is trained to reconstruct \( F_{ti} \) as much as possible. Therefore, the encoder network can maintain the informative knowledge. We utilize the reconstruction loss function \( L_{rec} \) to train our Paraphraser Module \( P_m \), which is formulated as:

\[
L_{rec} = \| F_{ti} - P_m(F_{ti}) \|^2
\]

(8)

Next, the trained encoder network \( P_{en} \) can be exploited to embed \( F_{ti} \), which can be written as:

\[
F'_{ti} = P_{en}(F_{ti})
\]

(9)

where \( F'_{ti} \) is the output of \( P_{en} \). Consequently, the construction of salience graphs of teacher networks could be modified as:

\[
\Gamma_{ti} = MP_{en}(F_{ti}) = MF'_{ti}
\]

(10)

**C. Adversarial Distillation**

As depicted in Fig. 2, we also adopt the adversarial training strategy [25] to transfer the high-order knowledge hidden in the outputs of the teacher network \( T \). In our Adversarial Distillation, the discriminator network \( D \) and the student network \( S \) are alternatively updated in each training iteration.

At first, the discriminator network \( D \) is trained to distinguish the input from \( S \) or \( T \) for a given input which consists of the concatenation of \( Y_s / Y_t \) and \( X \). The training of the discriminator \( D \) is constraint by adversarial loss function \( L_{adv} \), which is formulated as:

\[
L_{adv} = E_{Y_s \sim \mathcal{P}(Y_s)}(D([Y_s, X])) - E_{Y_t \sim \mathcal{P}(Y_t)}(D([Y_t, X]))
\]

(11)

where \([,]\) denotes the concatenation operation, and \( E(\cdot) \) represents the expectation operation. The \( L_{adv} \) is designed to calculate the wasserstein distance between \( T \) and \( S \) [20], [25]. Next, the student network \( S \) approximates the output distribution of the teacher network \( T \) as much as possible by maximizing \( L_{adv} \). As the discriminator \( D \) and \( S \) are alternatively trained, and the \( D \) is fixed in the training of \( S \), the \( L_{adv} \) for the optimization of \( S \) can be modified as:

\[
L_S = E_{Y_s \sim \mathcal{P}(Y_s)}(D([Y_s, X]))
\]

(12)

**D. Logits Distillation**

Furthermore, we introduce Logits Distillation [16] to enforce the student network \( S \) outputs more similar predictions with \( T \). For the logits of \( S \) or \( T \), we produce a soft probability distribution \( Z(\cdot) \) by exploiting a softmax layer, which can be written as:

\[
Z_{(\cdot)}(p)^m = \frac{exp(Y(\cdot)(p)^m/\tau)}{\sum_{k=1}^M exp(Y(\cdot)(p)^k/\tau)}
\]

(13)

where \( Z_{(\cdot)}(p)^m \) indicates the probability of \( m_{th} \) class of \( p_{th} \) pixel on \( Y(\cdot) \). The \( \tau \) is a temperature which is set to 1. Since it can work well with the setting of 1, we do not extra adjust the hyperparameter. The \( M \) is the total number of categories. The Kullback-Leibler divergence is employed to minimize the discrepancy between the teacher soft targets \( Z_t \) and the student soft targets \( Z_s \) as follows:

\[
L_{kd} = \frac{1}{H \times W} \sum_{p=1}^{H \times W} KL(Z_s(p)||Z_t(p))
\]

\[
= \frac{1}{H \times W} \sum_{p=1}^{H \times W} M \sum_{m=1}^{M} Z_s(p)^m log \frac{Z_s(p)^m}{Z_t(p)^m}
\]

(14)

**E. Semi-Supervised Learning**

Our method can improve the performance of compact networks without extra annotations compared with training from scratch. It could provide a novel semi-supervised paradigm by distilling knowledge from cumbersome networks. At first, we utilize annotated data to train our teacher network. Next, the student network is trained by both labeled and unlabeled data. Our student network is supervised under the groundtruth and our Graph Flow when the input images are annotated. On the contrary, our student network is only trained by our Graph Flow for unlabeled images.

**F. Objective Functions**

In the optimization of our proposed Graph Flow, the training procedures can be implemented through three steps. At first, we employ \( L_{rec} \) to train \( P_m \) while the parameters of pre-trained \( T \) are frozen. Next, the discriminator \( D \) is trained to discriminate \( Y_s \) and \( Y_t \) by minimizing \( L_{adv} \). At last, we optimize the student network \( S \) with an overall objective function \( L \) that contains the basic cross-entropy loss \( L_{ce} \), variation graph loss \( L_{vg} \), adversarial loss \( L_g \), and logits distillation loss \( L_{kd} \). The \( L \) can be presented as:

\[
L = L_{ce} + L_{og} - \lambda_3 L_g + \lambda_4 L_{kd}
\]

\[
= L_{ce} + \lambda_1 L_{vertex} + \lambda_2 L_{edge} - \lambda_3 L_g + \lambda_4 L_{kd}
\]

(15)
where the hyperparameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ are set to $10^{-5}$, $10^{-9}$, 0.1, 1 respectively.

As for our semi-supervised strategy for dual efficient medical image segmentation, our $\mathcal{L}$ can be modified as

$$
\mathcal{L} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{vg}} - \lambda_3 \mathcal{L}_g + \lambda_4 L_{kd}
$$

$$
\mathcal{L} = \mathcal{L}_{\text{cross}} + \lambda_1 L_{\text{vertex}} + \lambda_2 L_{\text{edge}} - \lambda_3 \mathcal{L}_g + \lambda_4 L_{kd}
$$

where labeled means that the cross-entropy loss is only used for labeled data.

IV. EXPERIMENTS

A. Datasets

1) Gastric Cancer Segmentation Dataset: The Gastric Cancer Segmentation Dataset [38] (Gastric Cancer) consists of 500 pathological images (400 images for training and 100 images for validation) cropped from whole pathological slides of gastric regions with the resolution of $2048 \times 2048$. The images in the dataset are manually annotated with two semantic categories: normal and cancerous area. Following our previous work [26], we crop each $2048 \times 2048$ image into four $1024 \times 1024$ patches.

2) Synapse Multi-organ Segmentation Dataset: The Synapse Multi-organ Segmentation Dataset [39] (Synapse) comes from MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, which contains 30 abdominal CT scans (18 CT scans for training and 12 CT scans for validation) with 3779 axial contrast-enhanced abdominal CT images. The volume sizes of CT scans are from $85 \times 512 \times 512$ to $198 \times 512 \times 512$. We employ the preprocessed version of Synapse Multi-organ Segmentation Dataset offered by [39] to conduct our experiments, which includes 8 abdominal organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, stomach.

3) Breast Ultrasound Images Dataset: The Breast Ultrasound Images Dataset [40] (BUSI) consists of 780 images with 437 benign cases, 210 benign cases, and 133 normal cases, which are collected from 600 female patients. Considering the main purpose of breast lesion segmentation in the clinical usage, we remove the normal cases that do not have breast lesion areas [10]. We random split 518 images for training and 129 images for validation. Since the images are different in their size, we resize the images with $256 \times 256$.

4) CVC-ClinicDB Dataset: CVC-ClinicDB [41] is a dataset containing 612 images which are extracted from 25 colonoscopy videos and selected 29 sequences from them. Following [11], [43], we also add Kvasir [42] with 900 images into our training set. Therefore, the total training set are extracted from CVC-ClinicDB and Kvasir containing 1450 images total. The 62 images from CVC-ClinicDB are used for validation. To unify the size of images, we resize the images with $256 \times 256$.

B. Implementation Details

We adopt the FANet [5] ($T_1$) and TransUnet [39] ($T_2$) as our teacher networks respectively, in which the FANet is an improved Unet based on self-attention mechanism and the TransUnet is a hybrid convolution and transformer architecture. Meanwhile, the lightweight Unet based on MobileNetv2 [44] (Mobile U-Net) is employed as our student network $S$. Moreover, we also adopt the lightweight networks, ENet [45] and ERFNet [46], as our student networks. The different teacher networks and student networks can examine the generalization of our method effectively. We use Adam as the optimizer of $T_1$, $S$, and $D$ respectively, with the weight decay $= 0.0002$. Following [39], the $T_2$ are optimized by SGD optimizer with learning rate 0.01, momentum 0.9 and weight decay 0.0001. For training $P_m$, we employ the standard SGD as optimizer with the momentum $= 0.9$, weight decay $= 0.0002$. The “ploy” learning rate policy $(1 - \frac{\text{iter}}{\maxiter})^{\text{power}}$ is introduced in the training of each networks, where the $\text{power}$ is set to 0.9. The initial learning rate is set to 0.003 for optimizing $S$, $T_1$, $P$ and 0.0002 for optimizing $D$, divided by 10 after every 50 epochs for $S$, $T_1$, $P$. In the training of overall networks, the epochs are set to 200 with the batch size of 8. Widely used data augmentation methods are also employed in our training [26]. All experiments are completed on the Pytorch platform with 2 NVIDIA TITAN XP GPUs. All experiments are repeated five times with different random seeds, and the results are presented in a format of means and standard deviations.

C. Evaluation Metrics

We employ different evaluation metrics to verify the performance of our method on four datasets due to their different modalities. (1) Gastric Cancer Segmentation Dataset: Following [38] and [26], we report the pixel accuracy (ACC) and the mean Intersection-over-Union (mIoU) of our results for evaluation. (2) Synapse Multi-organ Segmentation Dataset: Following [39], we employ the average Dice Similarity Coefficient (DSC) and the average Hausdorff Distance (HD) of the 8 abdominal organs. (3) Breast Ultrasound Images and CVC-ClinicDB Datasets: Following [10], [11], we employ the average Dice Similarity Coefficient (DSC) and the mean Intersection-over-Union (mIoU) of our results for evaluation.

Furthermore, We also compute the sum of point operations (FLOPs) and the number of network parameters (Params) to measure the model complexity. Besides, the non-parametric Wilcoxon test [47] is conducted to evaluate if improvements of our method are statistically significance with $p \leq 0.05$. At last, we also report the standard deviations among different samples (or patients) with the best-performance model in all experiments.

D. Results of Comparative Experiments

1) Results of the Comparisons with Other Knowledge Distillation Methods: We demonstrate the superiority of our method by comparing with the state-of-the-art popular knowledge distillation methods, as shown in Table. 1. Moreover, the visualization in Fig. 5 also shows the significant advantages of our method. Since KD [16], AT [48], FT [24] and VID [49] all focus on the mimicking of feature maps or the transformation of themselves from middle layer or outputs,
they exhibit similar performance and our method surpass them almost 3% in mIOU on Gastric Cancer. Compared with FSP [18], our method suppresses redundant activations except for patch-wise salience regions, which are unrelated to semantic-aware knowledge in deep layers, and focuses more on structural semantic-aware knowledge (class-wise spatial information and inter-class correlation) of an intermediate layer, which results to 2.05% increase in mIOU on Gastric Cancer. In addition, our method has significant reduction of HD on Synapse comparing with SKD [20] and IFVD [19]. Moreover, our method can improve 1.36% ACC on Gastric Cancer and 1.57% average DSC on Synapse compared with EMKD, which demonstrate that our method can perform better than EMKD [22] for medical image segmentation.

2) Results of the Comparisons with Our Prior Work:
To demonstrate the comprehensive extension of our Graph Flow with our preliminary CoCo DistillNet [26], we conduct quantitative and qualitative analysis, as shown in the last two rows of Table I and Fig. 6. Quantitatively, our method yields remarkable improvement compared to our previous work, especially in HD on Synapse with the 6.1235 mm reduction.

On Gastric Cancer, our method achieve the progresses of 0.9% on ACC and 1.44% on mIOU. Qualitatively, we can observe that the student network trained by Graph Flow is not influenced by the surrounding cancerous areas. Our student network could predict more correct normal areas (yellow boxes in second row left), and could distinguish different organs (red boxes in first row right). Moreover, the student trained by Graph Flow can discriminate left and right kidney (red boxes in the second row right) even without the discrepancy of spatial position (only a kidney exists in this slice). Furthermore, our method can help the student network avoid some error-prone details, as shown in the red boxes in second row left of Fig. 6. In particular, we exhibit the failed segmentation samples in last two rows of Fig. 6. Although all networks perform poorly on these samples, our method can correct some errors in details.

E. Results of Ablation Study

1) Results of the Patch-wise Salience Region Ablation: To report the effect of different sizes on salience regions \(P\), we conduct patch-wise salience region ablation on both Gastric Cancer and Synapse. As shown in Table II, the patch size \(9 \times 9\) can gain the best performance compared with other patch sizes on Gastric Cancer, which can improve the performance of \(S\) without Graph Flow Distillation by 2.63% (ACC) and 4.12% (mIOU) respectively. This can be explained by the large circular characteristics of the gastric cancer areas in pathological image. For Synapse, we observe that the patch size \(3 \times 3\) outperforms \(S\) without distillation by 4.98% (average DSC) and 13.7153 mm (average HD) respectively. This attributes to that Synapse has many small abdominal organs (e.g. aorta and gallbladder), and too large size of \(P\) causes the salience graph to obtain redundant and irrelevant class-wise information in certain channels when encoding small abdominal organs. In other words, the small abdominal organs are sensitive to the size of \(P\), but the large abdominal organs are relative insensitive. For example, as shown in Table II, the range of average DSCs of gallbladder can achieve 4.01% on different patch sizes. On the contrary, the range of average DSCs of liver only has 0.20%. In all our experiments, we select \(9 \times 9\) for Gastric Cancer, BUSI, and CVC-ClinicDB Dataset, and \(3 \times 3\) for Synapse.

2) Results of the Components Ablation: We conduct experiments to verify the contributions of each component in our
Graph Flow. To begin with, we demonstrate the positive effectiveness of our three different distillation methods as shown in Table. IV. To be specific, our Graph Flow Distillation can induce 4.98% improvement of average DSC on Synapse, and Adversarial distillation can promote 0.44% in mIOU for gastric cancer segmentation. Our overall Graph Flow can achieve a competitive performance compared with T1 (90.05% ACC vs. 88.69% ACC on Gastric Cancer, 79.50% average DSC vs. 78.70% average DSC on Synapse). In addition, we visualize the effectiveness of each component of our Graph Flow in Fig. 8.

Afterwards, the indispensability of our Paraphraser Module is verified in Fig. 7 and Table. V. Our Paraphraser Module can give 0.53% promotion on ACC for Gastric Cancer and 0.38% on average DSC for Synapse. Besides, we can observe that the performance degradation of our Graph Flow is more slight compared with our previous work when the Paraphraser Module is subtracted.

3) Results of the Generalization Ablation: As shown in Table. III, we leverage our Graph Flow on two different cumbersome networks: FANet and TransUnet, and three different lightweight networks: Mobile U-Net, ENet, and ERFNet. It is obvious that these lightweight networks have lower complexity than cumbersome network, while their lightweight designs are disadvantage for performance. However, our proposed knowledge distillation method can greatly improve the performance of such networks without any extra parameters. Especially, ENet, the most lightweight network, obtains only 1.56% lower than FANet in ACC for gastric cancer pathology image segmentation. Moreover, on CVC-ClinicDB, our method can improve the performance of ENet by 3.92% in average DSC and 3.41% in mIOU, when teacher networks is Transformer architecture. In addition, the visualization results of different student networks (ENet and ERFNet) with our method on BUSI and CVC-ClinicDB, which are taught by TransUnet, are shown in Fig. 9.

F. Results of Semi-Supervised Learning
At last, we explore the ability of our method in label-limited situations. For Gastric Cancer, we randomly select labeled data in the whole datasets according to several different labeled data proportion (20%, 40%, 60%, 80%, 100%). The Synapse is collected from CT scans from different patients. A CT scan of a certain patient could be included in both

### Table II

**Ablation Results of Graph Flow Based on Different Patch-Wise Salience Regions on Gastric Cancer and Synapse. We Also Report DSC of Every Abdominal Organ Except Average DSC and HD on Synapse.** 64 × 64 Refers to the Size of Feature Maps

| Methods       | Gastric Cancer | Synapse |
|---------------|----------------|---------|
|               | ACC  | mIOU  | ACC  | mIOU  |
| TI, FANet     | 0.906 | 0.819 | 0.790 | 0.625 |
| S Mobile U-Net| 0.884 | 0.748 | 0.731 | 0.427 |
| +Graph 4×4×64 | 0.879 | 0.776 | 0.782 | 0.450 |
| +Graph 3×3×64 | 0.879 | 0.775 | 0.782 | 0.450 |
| +Graph 5×5×64 | 0.871 | 0.781 | 0.779 | 0.450 |
| +Graph 7×7×64 | 0.874 | 0.777 | 0.773 | 0.450 |
| +Graph 9×9×64 | 0.867 | 0.766 | 0.772 | 0.450 |

Fig. 7. The loss curves of ablation analysis on Graph Flow without or with Paraphraser Module.

Fig. 8. The visualization of components ablation on Gastric Cancer and Synapse. **Zoom in** for better details.

Fig. 9. The visualization of generalization ablation on BUSI and CVC-ClinicDB. Teacher is TransUnet. **GF** indicates Graph Flow. **Zoom in** for better details.
labeled and unlabeled data, when we randomly select labeled data from the whole dataset according to these different labeled slices proportions. It would lead to the leak of the information of the unlabeled set. Therefore, for Synapse, we select a certain number of CT scans of patients as labeled data (2, 6, 10, 14, 18) and the rest of CT scans as unlabeled data. Then, We conduct 4 groups of experiments in each fixed proportion or numbers of labeled data: 1) training a teacher network only with labeled data; 2) training a student network only with labeled data; 3) training a student network only with labeled data by our Graph Flow (pre-trained teacher network from 1)); 4) training a student network with labeled and unlabeled data by our Graph Flow (pre-trained teacher network from 1)).

The results of semi-supervised learning on Gastric Cancer and Synapse are shown in Fig. 10. We also present the effectiveness visualization of semi-supervised learning in Fig. 11.

We can observe that the student network with labeled and unlabeled data in our GF framework outperforms the student network only with labeled data, and even surpass the teacher network (the number of labeled CT scans is 2 on Synapse). In addition, the superiority of our method with all dataset (labeled + unlabeled) is more noteworthy, when the proportion or numbers of labeled data is more limited (the gap between the yellow and green curve tends to expand with the reduction of labeled data).

V. DISCUSSIONS AND LIMITATIONS

In this section, we give detailed discussions about the experimental results and the specific application scenarios of our method. Furthermore, we discuss the method limitations.
A. Discussion on Different Experiments

1) Comparisons with Previous Distillation Methods: From the Table. I, we can observe that our method has statistically significant improvement compared with most state-of-the-art knowledge distillation methods. Specifically, our method has remarkable reduction of HD on Synapse. The lower HD means that the shape of our predictions are closer with the annotations. This indicates that our method can better represent holistic semantic-aware knowledge, which is crucial for medical image segmentation. Most of these methods in Table. I can only work well in one of the datasets. On the contrary, our Graph Flow exhibits prominent performance on the both datasets with different image modalities obtained from pathological slides and CT scans, respectively. We conjecture that these methods do not notice the knowledge hidden in the flow of cross-layers of teacher networks, which can teach student networks to extract more critical features for accurate organic or pathological segmentation. Moreover, the variation of semantic-aware knowledge between different layers is general for medical image segmentation networks. It indicates that our method has better generalization ability on different modalities.

2) Comparisons with Our Prior Work: From the last two rows of Table. I and Fig. 6, we can observe that our predictions are shape-closer with the annotations, as shown in the second row in Fig. 6. Besides, our method could eliminate extra error predictions better than our previous work as shown in Fig. 6. Therefore, our method can reduce the variances and biases of segmentation, which means it can produce relatively stable predictions even for complex images, as shown in last two rows Fig. 6. Moreover, we conjecture that our method can teach student to capture the spatial relations between different categories, because our salience graphs can encode class-wise spatial information. Moreover, our method also transfers the flow of inter-class correlation from teacher to student. Our Graph Flow can better cooperate with the supervision of ground truth than our prior CoCo DistillNet. On the other hand, our new method avoids the computing of Channel Mixed Spatial Similarity Module (CMSS) proposed by our previous work, which involves the resolution-wise multiplication operations ($B \times HW \times HW$) with the huge storage memory. Therefore, our Graph Flow does not add complexity and needs less storage memory than our previous work with a better performance.

3) Patch-Wise Salience Region Ablation: The different patch-wise salience regions mean that the salience graph has different sizes of respective fields on class-wise knowledge. As shown in Table. II, the comparisons between local and non-local ($64 \times 64$) salience regions demonstrate that our salience operation can eliminate redundant knowledge, thus the salience graph could focus more on favorable semantic class-aware knowledge. From Table. II, we can conclude that the dataset tends to select small patch-wise salience regions of $P$, if there are great differences in category size (e.g. Synapse). On the other hand, the dataset with resemblance in category size (e.g. Gastric Cancer) may employ relative large $P$.

4) Components Ablation: From Table. IV and Fig. 8, we could conclude that our Graph Flow predicts more correct semantic class-wise knowledge than student trained from scratch (the third column from last). Moreover, Adversarial distillation can complete the holistic alignment of the predictions between teacher and student (second row, penultimate column), while Logits Distillation is more beneficial for details (last column). As shown in Fig. 7, the Paraphraser Module can significantly stabilize training of our Graph Flow Distillation as it refines the knowledge of teacher by mitigating the gap between teacher and student. Besides, as shown in Table. V, the comparisons between Graph Flow and our previous work indicate that our proposed Graph Flow is more robust than our previous work. Furthermore, Graph Flow Distillation is more adapted with Paraphraser Module than our CoCo DistillNet.

5) Generalization Ablation: The generalization ability for different networks of knowledge distillation methods is crucial in medical image segmentation. The results of Table. III demonstrate that our method can extract knowledge from different cumbersome networks (conventional convolutional architecture or prevalent transformer architecture), and then transfer the knowledge to different lightweight networks. Furthermore, diverse lightweight networks with our Graph Flow can completely substitute cumbersome networks in resource-limited clinical diagnosis.

6) Semi-Supervised Learning: All of these shown in Fig. 10 and Fig. 11 confirm our method with unlabeled data can yield extra gains. It indicates our method can simultaneously solve resource-limited and annotation-limited issues, which is so called dual efficient (network efficient and annotation efficient) medical image segmentation.

B. Discussion on Application Scenarios

Medical image segmentation has achieved greatly progresses in the past several years, which benefits from the development of deep learning. However, these networks with outstanding performances always need huge computational cost and inference time. Therefore, they are impractical in some poverty-stricken rural hospitals of developing countries, especially in village or personal clinics. In these regions, extra purchasing of expensive computer-assisted diagnosis devices with GPUs may represent a heavy burden. Furthermore, too long inference time of these networks could damage the efficiency of real-time computer-assisted diagnosis (e.g. intra-operative stent segmentation [12]) for large hospitals in developed countries with tens of thousands of patients every day. Meanwhile, the lack of large-scale annotated medical datasets further hinders applications of high-performance networks in the real world.

Our Graph Flow can balance the trade-off between the performance and consumption. Our method can improve lightweight networks with comparative performances than cumbersome networks, without extra computation cost and inference time. Moreover, our method can achieve extra yields from extensive unlabeled medical images with a semi-supervised paradigm. Therefore, to achieve medical image
segmentation in the real world, our proposed Graph Flow could be employed.

C. Limitations

In this paper, we design a novel distillation method (i.e. Graph Flow Distillation) and propose a full-stage knowledge distillation framework combined with Adversarial Distillation and Logits Distillation. The distillation framework transfers the flow of semantic-aware knowledge, distills the knowledge of outputs, and utilizes the adversarial training on the prediction jointly. Our Graph Flow exhibits superior performance on different medical modalities, different cumber-some networks, and different lightweight networks. However, there are still some limitations in our method. Firstly, the performance gap between the student trained by our method and the teacher still exists, which is also a general limitation of knowledge distillation. Though in some cases, the student trained by our method can surpass the teacher when the annotated images are scarce, further research is needed to explore the lightweight student network with the teacher level performance. Secondly, our salience operation utilizes a fixed size of respective field for each channel. The fixed respective field may lead to the loss of important areas, since we tend to utilize small respective field when the medical images have great category-size differences. Though the results of patch-wise salience region ablation indicate that the category with large size may be insensitive for the size of respective field. If we could design more flexible respective sizes for different channels, our method will improve the student performance with better details. At last, the semi-supervised learning scheme and the knowledge distillation framework are embedded jointly. From the perspective of semi-supervised learning, our semi-supervised strategy may be deficient compared with the methods specifically designed for tackling the lack of annotated data. In addition, it is worth noting that our method is more applicable to harsh medical scenarios with data efficiency and network efficiency.

VI. Conclusion

In this paper, we propose a knowledge distillation method, Graph Flow, based on the feature flow of salience features from teacher network and student network. With the distillation of cross-layer graph flow knowledge and the outputs, the Graph Flow can significantly improve the performance of a lightweight network for dual efficient medical image segmentation framework. On four different-modality and multi-category datasets, we conduct a series of experiments to demonstrate the effectiveness of our method. Furthermore, our method provides a new paradigm for semi-supervised efficient medical image segmentation, which could jointly solve the scarceness of large-scale annotations and the resource-limited medical application problem.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., 2015, pp. 234–241.

[2] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis. (3DV), Oct. 2016, pp. 565–571.

[3] Z. Zhou et al., “UNet++: Redesigned skip connections to exploit multiscale features in image segmentation,” IEEE Trans. Med. Imag., vol. 39, no. 6, pp. 1856–1867, Jun. 2019.

[4] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “iU-Net: A self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, vol. 18, no. 2, pp. 203–211, Dec. 2020.

[5] 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 J. Biomed. Health Informat., vol. 25, no. 6, pp. 2071–2081, Jun. 2021.

[6] S. Takahama et al., “Multi-stage pathological image classification using semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 10702–10711.

[7] H. Tokunaga, Y. Teramoto, A. Yoshizawa, and R. Bise, “Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology,” in Proc. IEEE/CVF Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 12597–12606.

[8] Çiçek et al., “3D U-Net: Learning dense volumetric segmentation from sparse annotation,” in Proc. Int. Conf. Med. Image Comput.-Assist. Interent. (MICCAI), 2016, pp. 424–432.

[9] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, “Medical image segmentation using deep learning: A survey,” 2020, arXiv:2009.13120.

[10] C. Xue et al., “Global guidance network for breast lesion segmentation in ultrasound images,” Med. Image Anal., vol. 70, Jan. 2021, Art. no. 101989.

[11] T. Kim, H. Lee, and D. Kim, “UACANet: Uncertainty augmented context attention for polyp segmentation,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 2167–2175.

[12] Y.-J. Zhou et al., “Lightweight double attention-fused networks for intraoperative stent segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interent. (MICCAI), 2020, pp. 3–13.

[13] T. Lei, W. Zhou, Y. Zhang, R. Wang, H. Meng, and A. K. Nandi, “Lightweight V-Net for liver segmentation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 1379–1383.

[14] L. Zhang, L. Shi, J. C.-Y. Cheng, W. C.-W. Chu, and S. C.-H. Yu, “LPAPR-Net: Efficient vertebra segmentation from binocular wholspine radiographies,” IEEE J. Biomed. Health Informat., vol. 25, no. 7, pp. 2710–2721, Jul. 2021.

[15] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” Int. J. Comput. Vis., vol. 129, pp. 1789–1819, Mar. 2021.

[16] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.

[17] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “FitNets: Hints for thin deep nets,” 2014, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

[18] J. Yim, D. Joo, J. Bae, and J. Kim, “A gift from knowledge distillation,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Jun. 12, 2020, doi: 10.1109/TPAMI.2020.3001940.

[19] Y. Wen, L. Chen, S. Xi, Y. Deng, X. Tang, and C. Zhou, “Towards efficient medical image segmentation via boundary-guided knowledge distillation,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2021, pp. 1–6.

[20] D. Qin et al., “Efficient medical image segmentation based on knowledge distillation,” IEEE Trans. Med. Imag., vol. 40, no. 12, pp. 3820–3831, Dec. 2021.

[21] M. Q. Tran, T. Do, H. Tran, E. Tipputra, Q. D. Tran, and A. Nguyen, “Light-weight deformable registration using adversarial learning with distilling knowledge,” IEEE Trans. Med. Imag., vol. 41, no. 6, pp. 1443–1453, Jun. 2022.

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

[23] Z. Xu, Y.-C. Hsu, and J. Huang, “Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks,” 2017, arXiv:1709.00513.
[26] W. Zou, X. Qi, Z. Wu, Z. Wang, M. Sun, and C. Shan, “CoCo DistillNet: A cross-layer correlation distillation network for pathological gastric cancer segmentation,” in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Dec. 2021, pp. 1227–1234.

[27] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, “H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes,” IEEE Trans. Med. Imag., vol. 37, no. 12, pp. 2663–2674, Dec. 2017.

[28] J. Zhang, Y. Xie, P. Zhang, H. Chen, Y. Xia, and C. Shen, “Light-weight hybrid convolutional network for liver tumor segmentation,” in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 4271–4277.

[29] C. Guo, M. Szemenyei, Y. Yi, W. Wang, B. Chen, and C. Fan, “SA-U-Net: Spatial attention U-Net for retinal vessel segmentation,” in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 1236–1242.

[30] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, “Channel-wise knowledge distillation for dense prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 5311–5320.

[31] Y. Hou, Z. Ma, C. Liu, T.-W. Hui, and C. C. Loy, “Inter-region affinity distillation for road marking segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12486–12495.

[32] Y. Liu et al., “Knowledge distillation via instance relationship graph,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7096–7104.

[33] S. Zhou et al., “Distilling holistic knowledge with graph neural networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 10387–10396.

[34] S. Lee, “Graph-based knowledge distillation by multi-head attention network,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2019, pp. 141–154.

[35] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation,” Med. Image Anal., vol. 63, Jul. 2020, Art. no. 101693.

[36] L. Yu et al., “Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI), 2019, pp. 605–613.