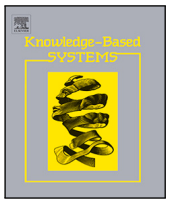
Knowledge-Based Systems 315 (2025) 113196

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Adaptive weighted multi-teacher distillation for efficient medical imaging segmentation with limited data

Eddardaa Ben Loussaief *, Hatem A. Rashwan, Mohammed Ayad, Adnan Khalid, Domemec Puig

Department of Computer Science and Mathematics of security, University Rovira I Virgili, Tarragona, 43007, Spain

ARTICLE INFO

Keywords:
Medical image segmentation
Lightweight student network
Multi-teacher distillation
Adaptive weighted distillation
Ensemble learning

ABSTRACT

Advances in deep learning models have significantly improved performance in medical tasks, but their complex structures and high computational requirements pose challenges for clinical implementation. Additionally, data privacy concerns limit the availability of comprehensive datasets needed to train accurate models. To address these issues, we propose a novel adaptive knowledge distillation (KD) framework for medical imaging segmentation that integrates intermediate and high-level feature pairwise relationships between teacher and student models. Our framework features adaptive multi-teacher distillation, where multiple teacher models, each trained on limited data from different sites and hospitals with various scanning protocols, distill their knowledge to a student model using adaptive weighting. This method allows each teacher to convey deep feature representations to the student's intermediate layers, enhancing performance without increasing complexity. To validate the efficacy of our framework, we conducted extensive experiments on two publicly available medical datasets, focusing on prostate and spleen tumor segmentation tasks. Our adaptive KD approach significantly improved dice scores by up to 9%, surpassing all tested baseline models. These results highlight the potential of our KD framework to enhance medical imaging segmentation while ensuring data privacy and security.

1. Introduction

Medical imaging modalities such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are extensively used for diagnostic purposes, particularly in segmentation and detection tasks that require pixel-level interpretation to produce segmentation masks. Despite their widespread use, image segmentation presents several challenges, including limited data resources and variations in image acquisition methods, such as different imaging modalities and scanning protocols. Deep learning (DL) networks, exemplified by models like UNet, [1], have demonstrated their efficiency in various medical tasks, especially semantic segmentation. However, these deep learning models are inherently complex and demand significant computational resources, challenging their deployment in real-time clinical settings.

To balance model performance with computational efficiency, lightweight models such as ESPNet [2] and ENet [3] have been developed. While these models reduce computational costs, they frequently suffer from degrading accuracy, creating a trade-off between efficiency and performance. Addressing this issue, knowledge distillation (KD)—a method introduced by Hinton et al. [4] has gained attention as a means to improve the performance of lightweight models. KD involves transferring knowledge from a larger, well-trained teacher model to a smaller student model, thus improving the student's performance while maintaining low computational demands.

Despite the growing success of KD in tasks like semantic segmentation by optimizing the similarity between teacher and student features [5] or by capturing feature pairwise relationships [6] to consider the spatially structured information, current methods still struggle with capturing intricate semantic details, particularly in the medical domain [7–9], where the spatial structure of data is critical. Most KD techniques rely on a single-teacher model, which limits the ability to distill diverse and complementary features to the student model. To address this gap, recent research has explored multi-teacher knowledge distillation, where the student model learns from multiple proficient teachers trained on distinct datasets, enhancing the generalization capability of the student model across diverse scanning protocols and institutions [10,11].

In the context of medical imaging, where privacy concerns and limited access to diverse datasets are prominent, multi-teacher KD offers a promising solution. Training models with data from multiple

* Corresponding author.
E-mail addresses: Eddardaa.benloussaief@urv.cat, eddardaa.benloussaief@estudiants.urv.cat (E.B. Loussaief).

https://doi.org/10.1016/j.knosys.2025.113196
Received 24 July 2024; Received in revised form 15 November 2024; Accepted 17 February 2025
Available online 3 March 2025

sources can improve generalization [12] but also raises concerns about biases introduced by differing scanning protocols. Additionally, the scarcity of data makes training high-performance teacher models even more challenging. Multi-teacher KD can help capture a broader range of features and patterns from diverse data sources, enhancing the student model's ability to generalize and adapt across different sources and protocols. Therefore, any distillation framework must be capable of training teacher models with limited data, potentially sourced from a single scanning resource. To date, no research has explored the potential of leveraging multi-teacher KD under these constraints, where each teacher is trained on limited data from a single source. Our approach aims to fill this gap by developing a KD framework that integrates multiple teachers trained on limited data to distill their knowledge a lightweight student model thus enabling improved segmentation performance and better generalization across different medical imaging datasets.

This work addresses the key challenge in medical imaging segmentation: how to improve the performance of lightweight models when training data is limited and diverse. We propose a multi-teacher knowledge distillation (KD) approach, where multiple teacher models, trained on limited and heterogeneous datasets, guide the student model collaboratively. Our adaptive multi-teacher distillation approach differs from traditional multi-teacher methods by dynamically weighting the influence of each teacher model based on the specific features and complexities of each input. Traditional multi-teacher distillation typically assigns equal or fixed weights to each teacher, which may not capture nuanced details in medical imaging data. By adapting weights per input, our method better handles diverse patterns and improves performance on complex medical imaging tasks, leading to enhanced accuracy and robustness. This approach aims to enhance the generalization and segmentation accuracy of the student model, overcoming the difficulties posed by variations in medical data and preserving patient privacy.

Our proposed multi-level learning scheme enables the student model to learn from the soft predictions of multiple teachers at a high distillation level, while also mimicking their feature representations at an intermediate level. This allows the student to incorporate a broader range of insights from diverse data sources without overfitting to a specific dataset. This addresses a crucial gap in the field, where most KD methods rely on a single teacher and are limited in their ability to handle complex, multi-modal medical data. We make the following key contributions:

- Adaptive Knowledge Distillation Framework: We introduce a novel adaptive knowledge distillation framework for medical imaging segmentation that combines intermediate and high-level feature transfer, enabling effective utilization of limited data while preserving data privacy.
- Mono-Teacher and Multi-Teacher Distillation Scenarios: Our approach supports both single-teacher and multi-teacher distillation scenarios. The single-teacher model provides a baseline, where the student learns from a single, well-trained teacher. In contrast, the multi-teacher model adaptively integrates knowledge from several teachers using a weighted approach, allowing the student to learn deep feature representations from diverse data sources. This adaptability is particularly critical when dealing with varied medical imaging protocols and datasets.
- Maintaining Model Simplicity: Despite these improvements, our approach enhances performance without introducing additional complexity to the student models, addressing the computational challenges associated with deploying deep learning models in clinical settings.
- Extensive Experimental Validation: We validate our proposed framework through extensive experiments on two publicly available medical datasets, focusing on prostate and spleen tumor segmentation tasks. The results show that our lightweight model achieves a significant improvement of approximately 9% in dice score compared to state-of-the-art methods, demonstrating the potential of our KD framework to enhance medical imaging segmentation while ensuring data privacy and security.

2. Literature review

2.1. Medical imaging segmentation

Significant progress in medial image segmentation for disease diagnosis has been achieved. Semantic segmentation falls into two categories: complex models for high-accuracy region segmentation and lightweight models for resource-limited devices, typically with lower accuracy.

Regarding complex models, Tian el.[13] proposed a PSPNet architecture by integrating fully connected networks for MRI semantic segmentation. Liu et al. [14] proposed an FCN with ResNet50 as the backbone for detecting the prostate zones with the mechanism of feature pyramid attention. Zhu et al. [15] introduced a novel domain adaptation-based method, presenting a boundary-weighted segmentation to exceed the limited source data. To deal with the underfitting problem during the training, Liu et al. [16] implemented a robust multi-site segmentation model by aggregating prostate MRI from multiple sources. In [17], the authors developed an automatic segmentation tool using a fully connected conditional random field that works on both 2D images and 3D medical volumes without any prior information or training process. Moon et al. [18] developed an end-to-end automated pipeline for abdominal spleen segmentation using Resnet as a backbone. This pipeline allows the users to perform the segmentation into 3 stages, i.e. preprocessing the input data, segmentation with deep learning methods, and 3D visualization of the generated masks. In [19], Mihaylova et al. proposed an automated spleen segmentation in MRI sequence, they used a template matching method to set the initial active contour at which the segmentation starts. Their approach involves splitting the MRI sequence into two parts where the middle slice represents the border. Based on the middle image's segmentation, the next slice's mask segmentation is created through a matching technique. Huo et al. [20] developed an image-to-image conditional generative adversarial network (cGAN) to address the spatial variations of the large spleens in multi-modal MRI scans. In most of the aforementioned DL models, semantic medical segmentation has several drawbacks: highly accurate models for semantic segmentation tend to be large and complex, making them difficult to deploy on devices with limited computational resources. Then, complex models often require significant computational power and time for inference, which can be impractical for real-time applications. They are prone to overfitting, especially when trained on small datasets. Large models can be difficult to interpret, making it challenging to understand their decision-making process. Highly accurate models for semantic segmentation tend to be large and complex, making them difficult to deploy on devices with limited computational resources.

Thus, several studies have explored the use of lightweight models for medical imaging segmentation. In one study [21], the authors tested Unet, ENet, and ERFNet for segmenting prostate MRI images, finding that Unet achieved the highest Dice score, followed by ENet. The performance of ENet (dice of 0.835) here was attributed to the limited amount of training data. Ma et al. [22] developed a method for segmenting Tooth CT images using an ENet model enhanced with an attention mechanism, achieving a Dice score of 0.85. Additionally, Nicolas et al. [23] employed the 3D ESPNet model to segment brain tumors from 3D MRI volumes, reaching an accuracy of 0.785. However lightweight models face several challenges: they typically achieve lower accuracy compared to more complex models because they have fewer parameters and thus less capacity to capture intricate patterns in the data. Their simplified structure can result in inadequate generalization across different datasets or data variations. Their limited capacity can

hinder their ability to capture complex structures in medical images (e.g., MRI), leading to poorer segmentation quality. Then, they are more sensitive to noise and variations in the data, which can negatively impact their performance.

To address the drawbacks of lightweight models, the concept of Knowledge Distillation (KD) has emerged [4]. KD allows a lightweight model to be trained under the guidance of a more complex model, enabling it to mimic its performance while maintaining a lower number of parameters and reduced computational resource requirements.

The literature highlights considerable advancements in both complex and lightweight medical segmentation models. However, a notable research gap persists in effectively integrating knowledge distillation (KD) with advanced segmentation techniques to balance accuracy and computational efficiency. Existing studies fall short of providing comprehensive frameworks that leverage multi-teacher distillation to optimize lightweight models for specific medical imaging tasks. To address this gap, this paper proposes an optimized KD framework designed for lightweight segmentation of MRI prostate and CT spleen scans. Our approach enables the student model to simultaneously learn from multiple teacher models, enhancing its performance while maintaining computational efficiency.

2.2. Knowledge distillation for medical imaging

KD allows a lightweight student model to learn from the more accurate predictions and feature representations of a larger teacher model [4]. This process helps the student model achieve higher accuracy than it would have on its own, bridging the performance gap between lightweight and complex models.

Recently, researchers have witnessed the use of knowledge distillation in medical imaging segmentation. For instance, Wang et al. [24] have proposed an efficient self-supervised KD framework that encodes the intra-instance feature representation for skin lesion classification. In [25,26], a mutual distillation has been proposed that adopts a cross-modality KD framework for cardiac and abdominal organ segmentation. In [27], an efficient distillation scheme for brain segmentation has been proposed, it is based on coordinate distillation that incorporates space and channel information. In [7], to boost the lightweight model to mimic the middle features extracted from the teacher network during the training, Qin et al. [7] have constructed a novel module i.e. for liver and kidney segmentation that guides the student network by re-scaling the features' student to imitate the features maps of the teacher. Noothout et al. [28] have adopted two ensembles of convolutional neural networks as teacher models to segment the brain MRI imaging, chest CT, and cardiac cine-MRI. Xu et al. [29] have proposed a growing teacher assistant network (GTAN) along with the main teacher to optimize the disparity of teacher and student model sizes. Zhao et al. [8] have proposed a structured distillation framework that investigates a region graph distillation to exploit the higher-order representational capabilities of graphs to enable the student network to mimic structured information from the teacher better. In [30], the authors have introduced a Prototype Knowledge Distillation (ProtoKD) to improve medical image segmentation when only single-modality data is available, addressing the common issue of missing modalities in clinical practice. ProtoKD transfers knowledge from a multi-modality trained teacher to a single-modality student model by distilling pixel-wise knowledge and intra- and inter-class feature variations.

In summary, the literature on KD in medical imaging demonstrates steady progress; however, a significant gap persists in the adoption of multi-teacher distillation strategies. Most existing approaches rely on a single-teacher model to train student models, which limits the ability to fully exploit diverse and complementary knowledge from multiple high-performing teachers. This limitation is particularly pronounced when working with limited medical datasets, where multi-teacher KD could enhance the generalization and robustness of lightweight models. To address this gap, we designed a distillation pipeline that incorporates multi-teacher KD, a strategy that remains underexplored in medical image segmentation. This emerging approach shows immense promise for advancing the field. Specifically, this work introduces an innovative adaptive and ensemble KD strategy aimed at significantly improving segmentation performance while maintaining computational efficiency. Furthermore, medical imaging researchers have consistently highlighted the strong segmentation capabilities of models such as PSPNet [13] and Deeplab [31]. Building on their proven effectiveness, we utilize these models as teacher networks within our framework to maximize knowledge transfer to the student model.

3. Methodology

3.1. Overview

As shown in Fig. 1, we present a refined distillation pipeline featuring two pivotal models: a high-performance teacher model and a lightweight student network optimized for fast segmentation. Both models operate on 3D imaging data (MRI or CT) arranged as a stack of 2D slices, producing predictions for each slice. As depicted in Fig. 1, our structured distillation framework encompasses two main learning approaches to maximize robustness and adaptability.

Firstly, we employ mono-teacher distillation, wherein the student gains knowledge from a single, proficient teacher who has been extensively trained. The second scenario involves multi-teacher distillation, where the student network is trained by aggregating knowledge from multiple teachers, each trained on a distinct subset of the data. By integrating insights from multiple teachers, the student model becomes more resilient to noise and biases associated with individual teachers and can better generalize across diverse inputs and conditions. Averaging predictions from multiple teachers helps reduce overfitting by promoting generalizable and adaptive patterns instead of memorizing specific examples. This approach is particularly advantageous for medical imaging tasks, as the combined knowledge from multiple sources enhances the model's effectiveness and efficiency over traditional mono-teacher distillation approaches.

Thus, our primary objective is to equip the student with the capability to acquire additional insights from various teachers, enabling it to dynamically adapt to real-time medical imaging segmentation across diverse hospital datasets while addressing concerns related to the sharing of private data. To demonstrate the efficiency of our distillation pipeline, we adopt the publicly available MRI prostate [16] and CT Spleen [32,33].

For the mono-teacher transferring, we used the full data, while for multi-teacher distillation, we divided the larger dataset into smaller subsets (two subsets for the Spleen dataset and three for the Prostate dataset) and trained each teacher separately on its respective subset. This method enables the ensemble of teachers to distill the student network through a unified transfer process, applying consistent losses across both mono- and multi-teacher distillation. In this paper, we conducted a multi-level knowledge transfer approach. Our methodology encompasses feature-based distillation facilitated by attention transfer, inspired by [34], and logit-based distillation utilizing the Kullback–Leibler (KL) divergence to align teacher and student predictions. In line with the multi-teacher distillation approach, we employed dual teacher distillation for spleen data and trio-teacher for prostate MRI to enhance our experimental setup and increase model diversity, considering the data sources available for training the teacher models.

Our framework is designed to scale efficiently with larger datasets by training multiple teacher models independently on different data subsets or imaging modalities. This approach supports parallel processing, making it feasible to handle high data volumes. Additionally, the adaptive weighting process is modality-agnostic, allowing application across various medical imaging types beyond MRI and CT. By dynamically adjusting the weighting mechanism to reflect diverse data

distributions, our framework generalizes effectively across different imaging modalities and datasets acquired under varying conditions, highlighting its potential for broader applications in medical imaging. For our experiments, we utilized publicly available prostate MRI and spleen CT datasets, chosen for their diversity in data sources. However, our framework is readily adaptable to any multi-source medical imaging data.

Our approach guarantees that the teacher models cover a broad spectrum of data variations without introducing redundancy that could cause overfitting issues. Furthermore, we ensure that the student model effectively learns from the ensemble of teacher models, which might necessitate specific training methods utilizing efficient distillation loss functions. To achieve that, in our methodology, we follow:

- Carefully select teacher models that cover a broad spectrum of data variations, such as PSPNet [13] and Deeplab [31]. This can be done by choosing teacher models trained on diverse subsets of the data, using different architectures. Ensuring diversity among teacher models helps prevent redundancy and ensures comprehensive coverage of the dataset by:
- Using ensemble learning based on weighted averaging to combine predictions from multiple teacher models, capture a broader range of knowledge, and reduce the risk of overfitting.
- Implementing regularization techniques to prevent overfitting when training the student model on the ensemble of teacher predictions. Regularization methods such as dropout, weight decay, or early stopping can help the student model generalize better and avoid memorizing noise present in the teacher models.
- Designing and utilizing efficient distillation loss functions tailored to the ensemble of teacher models that could be trained with limited data. These loss functions should encourage the student model to learn from the ensemble's collective knowledge while balancing between different teachers' contributions. Techniques such as knowledge distillation, where the student model learns to mimic the soft labels or intermediate representations of the teacher models, can be particularly effective in this context.
- Developing a training procedure that effectively integrates the knowledge from multiple teacher models into the student model's learning process. This involves iterative training steps where the student model is trained using both the original dataset and the predictions from the teacher models, updating its parameters to minimize the distillation loss. It is crucial to maintain a balance between learning from the ensemble of teachers and preserving the student model's ability to adapt to new data.

3.2. Standard mono-teacher knowledge distillation

This work incorporates a dual-level knowledge transfer strategy: an intermediate transfer for feature-based distillation and a high-level transfer for prediction-based distillation. To begin, Logits-based distillation has been conducted in [4] as an efficient architecture to learn the student network from the soft targets provided by a well-optimized teacher model. In our work, we refine this approach by proposing a distillation architecture that utilizes soft targets comprising class probabilities and essential information crucial for the student model to emulate the output of the pre-trained teacher. The transfer probability is as follows:

$$p_i = \frac{exp(z_i/\lambda)}{\sum_j exp(z_j/\lambda)}, \tag{1}$$

where p_i indicates the input probability and z_i denotes the ith class of the logits output, i.e., z_0 background and z_1 prostate. λ is a hyperparameter called the temperature to balance the distillation loss. Inspired by [4], we determine a prediction distillation that enables the lightweight model to learn the prediction probability of the final output segmentation of the teacher model, i.e., softmax output. The logits-based distillation loss is precisely measured by Kullback–Leibler (KL) divergence across the student's and teacher's softmax predictions. Thus, prediction distillation loss is defined by:

$$KL_{loss} = \frac{1}{N}\sum_{i}^{N} KL(p_i^s \| p_i^t), \tag{2}$$

where the KL divergence function is denoted by $KL(.)$, where $N = w \times h$, all the pixels in segmentation, and p_i^s and p_i^t indicate the probability of ith pixel pair in the segmentation map of the student's and teacher's softmax outcome, respectively.

Along with the prediction distillation loss measured in (2), we explore an intermediate distillation loss to conduct the feature-based distillation. We follow the attention transfer in [34] that considers the absolute value of a neuron's activation. Specifically, we analyze the feature maps f, structured as $C \times w \times h$, by summing the absolute values along the channel dimension C to derive an attention value of the original features f. We denote at_i^s and at_j^t the attention feature maps extracted from the ith and jth layers for the student and teacher networks respectively. Thus, the attention transfer's loss can be calculated by:

$$AT_{loss} = \sum_{i,j} \| \frac{at_i^s - at_j^t}{\|at_i^s - at_j^t\|_2} \|_1, \tag{3}$$

where $\|.\|_1$ and $\|.\|_2$ are the L1 and L2 normalization. Furthermore, in [7], the authors introduced a method based on region contrast to identify the tumor boundary within the feature maps by utilizing the labeled segmentation mask. They assessed region contrast by comparing the similarity between the region information derived from the student and teacher independently. We utilize this region-based approach to blend the feature maps of the student and teacher and enhance the similarity between them through optimization. As such, we employ the final feature layer of each teacher's encoder and decoder sections to guide knowledge distillation. Given the intermediate distillation loss across feature maps for both teacher and student networks, we employ a loss function that combines the attention-based loss and the region contrast-based loss. This loss is defined as follows:

$$Mid_{loss} = AT_{loss} + \sum_{i,j} \|rc_i^s - rc_j^t\|_2, \tag{4}$$

where rc_i^s and rc_j^t represent the region contrast vectors of the ith and jth layers for the student and teacher, respectively (as described in [5,7]). The attention loss can help the student model to learn to focus on the most relevant parts of the teacher's feature maps. Region contrast involves transferring relational information across networks, allowing the student model to emulate the contrast between foreground and background regions. This process enables the student model to discern and mimic the distinctive features in foreground and background areas. To assess the accuracy of the student's segmentation mask against the labeled segmentation mask, we utilize the segmentation loss Seg_{loss}. This loss function incorporates the Lovasz-Softmax loss [35], known for its effectiveness in addressing class imbalance and ensuring boundary alignment. In addition to the Lovasz-Softmax loss, we incorporate the dice similarity loss [32] to further refine the segmentation evaluation.

$$Seg_{loss} = \alpha_1 Dice_{loss} + \alpha_2 Lovasz_{loss}. \tag{5}$$

where α_1 and α_2 are hyper-parameters set to 0.2 and 0.3, respectively.

As shown in Fig. 1, to obtain an end-to-end trainable student network, we utilize a blend of losses to execute the global distillation loss, referred to as KD_{loss}. Therefore, the complete distillation process for the single-teacher scenario is defined by the subsequent formula.

$$KD_{loss} = Seg_{loss} + \alpha Mid_{loss} + \beta KL_{loss}, \tag{6}$$

α and β are set to 0.1. The hyperparameters setting was settled after empirical attempts to conduct efficient segmentation results.

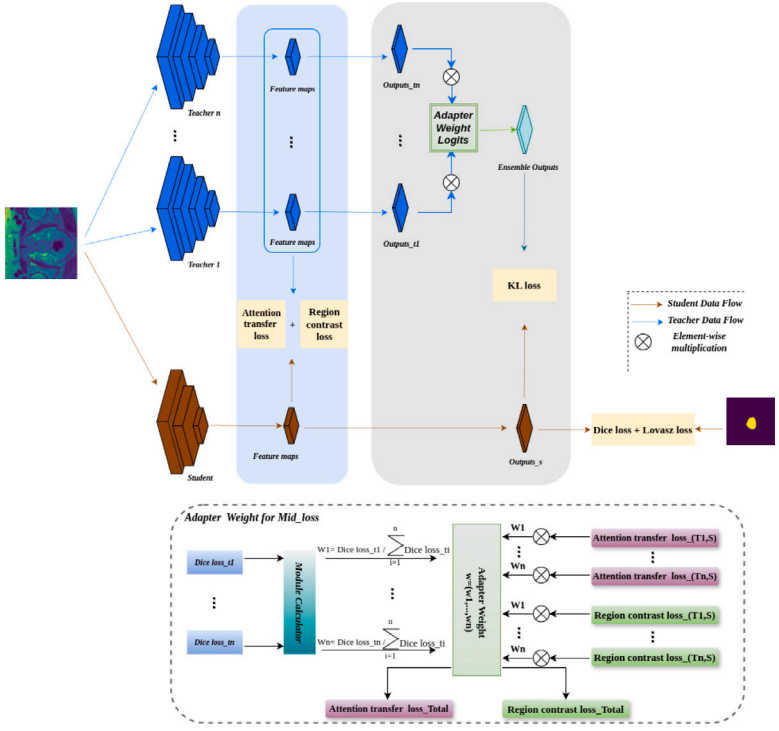


Fig. 1. The overall framework of our multi-teacher distillation method.

3.3. Adaptive ensemble learning

Single-teacher distillation has the potential to enhance the performance of lightweight models significantly. However, leveraging the ability to learn from multiple resources simultaneously and equally can empower the student model to capture a broader range of characteristics from various teacher networks. Recent research has tested multi-teacher knowledge as a potential way to enable the student model to learn from well-performed teachers with different complex architecture. Previous research on multi-teacher distillation learning has primarily focused on the aggregation technique that combines the outputs of multiple teachers into a single teacher-like output by averaging the predictions (or logits) of all teachers, which is then used to train the student model [36]. Moreover, Weighted Multi-Teacher Distillation [37] has been used to assign different constant weights to each teacher based on their performance or relevance to the specific task instead of equally considering all teachers. Then, the Selective Multi-Teacher Distillation is where the student is trained by selecting or voting for the best teacher (or a subset of teachers) for each specific instance or domain [38,39]. All the multi-teacher distillation methods mentioned previously have several drawbacks, i.e. Averaging the teachers' outputs might dilute individual expertise, leading to a loss of critical, nuanced knowledge. In addition, selecting or tuning the weights can be non-trivial and may require additional hyperparameter tuning or a sophisticated weighting mechanism. In contrast, our work introduces a new Adaptive Multi-Teacher knowledge distillation framework to adjust the contributions of each teacher model dynamically, improving robustness and generalization across various medical imaging domains. This framework can dynamically assign varying weights to different teacher models based on the complexity and characteristics of the input data. In medical imaging tasks, especially when dealing with heterogeneous datasets. Here, we innovate by dividing the available heterogeneous data considering its sources into subsets, allowing us to train each teacher network separately. This approach employs an adaptive strategy that enables students to learn from all teachers simultaneously, utilizing adaptive weights to compute the fusion of teacher models. Therefore, we evaluate the teachers' predictions using the following formula, $p_i^t = w_j \sum_j p_j^t$, where $j = 1 : n$ and n represents the total number of teachers. If we have n teacher models, then the set of adaptive weights obtained from these teachers is denoted as $w_j \in \{w_1, w_2, \ldots, w_n\}$. Therefore, the calculation of the adaptive weight w_j is determined by:

$$w_1 = \frac{Dice_{loss}(p^{t_1}, y)}{\sum_j (Dice_{loss}(p^{t_j}, y))}$$

$$w_2 = \frac{Dice_{loss}(p^{t_2}, y)}{\sum_j (Dice_{loss}(p^{t_j}, y))}$$

$$\vdots$$

$$w_n = \frac{Dice_{loss}(p^{t_n}, y)}{\sum_j (Dice_{loss}(p^{t_j}, y))}$$

where y is the ground-truth mask, considering multi-teacher scenario, and the student's output p_i^s using the formula given in Eq. (2), where p_i^t is defined as the weighted sum of the individual teacher outputs p_j^t, $p_i^t = w_j \sum_j p^{t_j}$.

Additionally, we compute the Mid_{loss} across the teachers and the student network using $w_j \in \{w_1, w_2, \ldots, w_n\}$ given above as shown in the following formula:

$$Mid_{loss} = w_1 Mid_{loss}(t_1, s) + w_2 Mid_{loss}(t_2, s) + \cdots + w_n Mid_{loss}(t_n, s). \quad (7)$$

Hence, the overall loss for the multi-teacher knowledge distillation framework will be computed using Eq. (6). We conduct our segmentation task on MRI prostate and CT spleen datasets, with details of the data collection provided in the next section. However, our adaptive multi-teacher distillation approach is not limited to these specific modalities or organs; it has the potential to perform effectively on a wide range of 3D medical scans or simpler modalities using the same process.

The selection of these datasets is based on their heterogeneity and availability. We expect that as the input data increases in size and diversity, our framework will demonstrate even stronger performance, as it is designed to adapt and handle complex and varied data more effectively.

Table 1
Total number of training data w/o augmentation for prostate and spleen segmentation tasks.

Strategy	Singe-teacher	Multi-teacher		
		T1	T2	T3
Prostate	2844	1095	1113	638
Spleen	6848	3829	3019	–

4. Experiments

4.1. Datasets

MRI Prostate : We conducted our experiments on a publicly available MRI prostate dataset [16], where the data is collected from three various institutions, PROMISE12 [40], NCI-ISBI2013 [41], and I2CVB [42]. The data consists of 116 patient cases with a total slice number of 1740 with their corresponding segmentation masks.

CT Spleen : Spleen data is sourced from Decathlon [32] and Duke [33] datasets. The Duke Spleen Data Set (DSDS) [33] comprises in total 109 CT and MRI volumes (6322 images) from 69 patients with chronic liver disease and portal hypertension. For our work, we focused on 69 axial CT volumes stored in DICOM format, which underwent preprocessing to convert the DICOM files for each patient into NIfTI files. This conversion facilitated seamless integration with the decathlon dataset. Decathlon spleen [32] includes a total of 61 CT volumes, with 41 volumes allocated for training purposes and 20 volumes for testing.

In our approach to mono-teacher distillation, we divided the data into training and testing sets, with an 80% to 20% ratio allocation, respectively. In the case of multi-teacher distillation, we employed three distinct teachers for the MRI prostate. To facilitate this, we partitioned the dataset into three subsets equally in terms of the number of samples. We have a total of 2846 and 1045 slices for the training and testing respectively (see Table 1), allowing us to train each teacher individually with a limited number of images. For the spleen data, we implemented a dual-teacher framework. The first teacher was trained using the Decathlon dataset, while the second teacher performed on the Duke dataset.

We applied a comprehensive suite of data augmentation techniques for the training and validation sets to increase sample diversity and improve model robustness. These augmentations include:

- Random Rotation with an angle limit of ±90 degrees.
- Random Scaling range of ±20%, and Shift with limit of 0.0625.
- Random Horizontal and vertical flips with a probability of 0.5.
- Random elastic transformations, grid distortions, and optical distortions with probabilities of 0.1.
- Random Brightness and Contrast with a probability of 0.2.
- Coarse Dropout which randomly masks small rectangular regions in the images with a probability of 0.2.

We used the Albumentations library to implement the described augmentation techniques, which proved effective in enhancing the model's generalization, particularly on challenging boundary regions. This approach significantly contributed to the model's overall robustness and performance consistency.

4.2. Evaluation metrics

In the scope of medical image segmentation, the Dice similarity coefficient (Dice) stands out as a commonly employed metric to evaluate the performance of each method. The Dice coefficient serves to compute the overlapping between two sets, offering a numerical representation typically ranging from 0 to 1. A greater Dice coefficient denotes a stronger similarity between the segmented area and the ground truth, indicating higher accuracy in the segmentation process. In Addition, the Dice loss enhances model performance, particularly in handling imbalanced datasets and ensuring better delineation of critical shapes. It is expressed as:

$$DICE(P,G) = \frac{2|P \cap G|}{|P| + |G|}, \tag{8}$$

where P and G denote the prediction and the ground truth of the tumor mask, respectively.

Additionally, we provide the volumetric overlap error (VOE) as an extra evaluation metric, which measures the prediction's error rate. VOE is defined as:

$$VOE(P,G) = 1 - \frac{|P \cap G|}{|P| + |G|}. \tag{9}$$

It is important to note that VOE differs from the Dice coefficient in their interpretation. While a higher Dice coefficient implies better network performance, VOE serves as an error metric, where smaller values (or their absolute values) are desired.

Most distillation methods in medical imaging commonly use the Dice similarity coefficient and Volume Overlap Error (VOE) to evaluate performance and benchmark against state-of-the-art (SOTA) approaches. In alignment with this standard practice, we have selected these two metrics as the primary evaluation criteria for our framework. These metrics not only enable us to assess the effectiveness of our approach but also facilitate a meaningful comparison with other knowledge distillation (KD) methods in the field of medical imaging.

4.3. Setup

We conducted a comprehensive series of experiments to assess the efficacy of our distillation approach. For our teacher networks, we opted for robust complex off-the-shelf segmentation models, such as DeepLabV3+ [31], PSPNet [13], FCN_ResNet101 [14], and ResNeXt-101-32x8d [43]. Regarding the student models, we selected three lightweight networks that are used for medical image segmentation, namely ENet [3], ESPNet [2], and MobileNetV2 [44]. Table 2 presents a comparison between teacher and student networks in terms of the number of parameters and FLOPs, highlighting the significant differences between the complex and lightweight models.

All tested networks were trained and evaluated using the official PyTorch setup on a GeForce RTX 3080 ti (16 GB). During training, we employed the Adam optimizer with default configuration ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and initialized the learning rate at 0.01. Additionally, we utilized CyclicLR to adjust the learning rate, setting the minimum learning rate to 0.000001 and a step size of 2000. The training was conducted until convergence over 100 epochs.

4.4. Results

4.4.1. Mono-teacher distillation

Our objective is to train a student network using a single teacher in the initial phase of our distillation process. The teacher network has undergone thorough training on datasets related to prostate and spleen tumors. To demonstrate the effectiveness of our distillation approach in improving the efficiency of a lightweight model, we conducted an empirical series, where, we employed ENet [3], ESPNet [2], and MobileNetV2 [44] as student networks, all possessing a low parameter count. Conversely, we selected sophisticated teacher networks, namely DeepLabV3+ [31], PSPNet [13], FCN_ResNet101 [14], and ResNeXt-101-32x8d [45] to guide the student models as mentioned previously. For DeepLabV3+ and PSPNet networks, we utilized ResNet50 as a backbone, while for FCN, we employed ResNet101.

These networks have been specifically designed for medical imaging tasks and are known for their high accuracy in state-of-the-art segmentation methods. Figs. 2 and 3 illustrate the improvement in dice scores across different student models, particularly emphasizing the

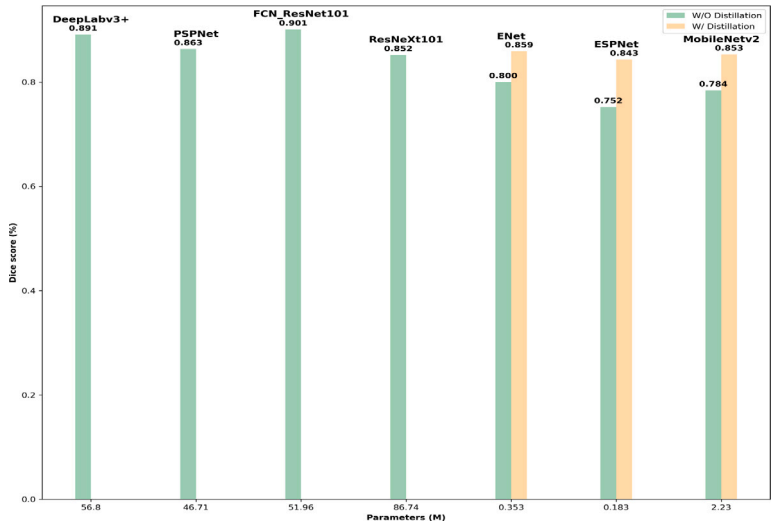


Fig. 2. Effectiveness of Distillation on Model Dice Score for prostate segmentation. We adopt ENet, ESPNet, and MobileNet as student models, each distilled from one of the following teacher networks: PSPNet, FCN_ResNet101, DeepLabV3+, and ResNeXt101. The green bars indicate the dice score achieved with baseline training, whereas the purple bars show the student models' dice scores after distillation from the teacher networks.

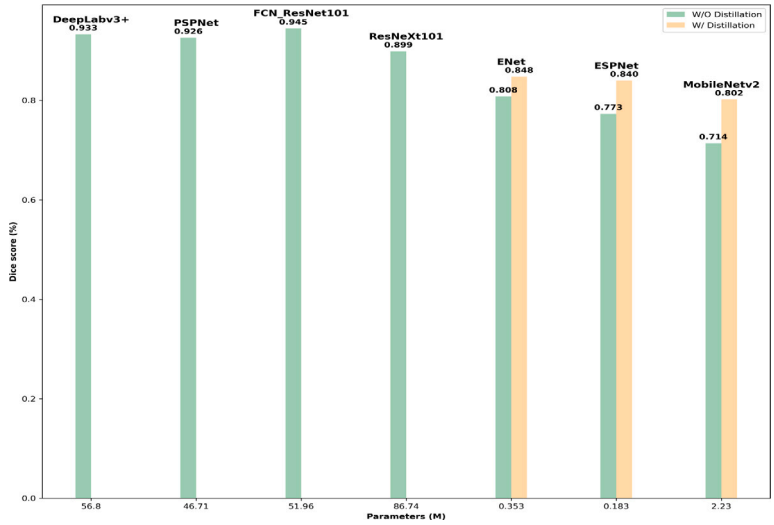


Fig. 3. Effectiveness of Distillation on Model Dice Score for spleen segmentation. We adopt ENet, ESPNet, and MobileNet as student models, each distilled from one of the following teacher networks: PSPNet, FCN_ResNet101, DeepLabV3+, and ResNeXt101. The green bars indicate the dice score achieved with baseline training, whereas the purple bars show the student models' dice scores after distillation from the teacher networks.

enhancement achieved with fewer parameters. Now, let us delve into the quantitative segmentation results for MRI prostate and CT spleen.

MRI Prostate segmentation: Table 3 demonstrates that, with the appropriate selection of the teacher networks for prostate MRI segmentation, all students are adept at learning and mimicking the teachers' capabilities and achieving higher performance, surpassing the performance of the teachers themselves. It is obvious to remark that ENet, ESPNet, and MobileNetV2 embrace the maximal improvement of 5.87% (80.03 to 85.9), 9.1% (75.5 to 84.3) and 6.9% (78.4 to 85.3) in dice score, respectively. Although all student networks show enhancement, they still fall short of the teacher networks' performance, albeit achieving efficiency with fewer parameters. The most notable improvement was achieved with ESPNet with FCN_ResNet101 as a teacher where the Dice score increased by 9.1% from a baseline of 75.2% to 84.3% with distillation.

CT Spleen segmentation: Table 3 illustrates the outcomes of employing our mono-teacher distillation framework for CT Spleen segmentation. We conducted practical experiments utilizing the same array of models for both teachers and students. The teacher networks attained a promising dice score with the Spleen data compared to the baseline training with Prostate data. This initial improvement can be attributed to the inherent characteristics of the data; specifically, CT scans for the Spleen organ exhibit greater intensity and clearer contrast compared to MRI scans for the prostate organ. Furthermore, from a quantitative perspective, the amalgamation of Duke and Decathlon datasets provides a rich array of training samples to train the teacher models

Table 2
The rank of the models in ascending Order of the number of parameters.

Method	Params(M)	Flops(G)
ESPNet	0.183	1.27
ENet	0.353	2.2
MobileNetV2	2.23	19.84
PSPNet	46.71	207.7
FCN_ResNet101	51.96	199.74
DeepLabV3+	56.8	273.94
ResNeXt101	86.74	96.54

Table 3
Cross experiments results between a single teacher and student on prostate and spleen data. "w/o" denotes the baseline performance, and "w/" stands for the performance with our distillation method.

Teacher networks	Prostate		Spleen	
	Dice	VOE	Dice	VOE
T1: DeepLabV3+	0.891	0.101	0.933	0.046
T2: FCN_ResNet101	0.901	0.085	0.945	0.039
T3:PSPNet	0.863	0.107	0.926	0.052
T4:ResNeXt101	0.852	0.143	0.899	0.099
Student Networks and their performances distilled from different teachers				
ENet: w/o	0.800	0.228	0.808	0.217
T1: w/	0.846	0.163	0.834	0.198
T2: w/	0.859	0.113	0.848	0.129
T3: w/	0.842	0.147	0.825	0.107
T4: w/	0.837	0.192	0.824	0.113
ESPNet: w/o	0.752	0.412	0.773	0.388
T1: w/	0.819	0.153	0.804	0.118
T2: w/	0.843	0.129	0.840	0.102
T3: w/	0.837	0.130	0.807	0.104
T4: w/	0.805	0.169	0.798	0.353
MobileNetV2: w/o	0.784	0.310	0.714	0.536
T1: w/	0.842	0.127	0.761	0.235
T2: w/	0.838	0.130	0.802	0.182
T3: w/	0.853	0.104	0.752	0.374
T4: w/	0.801	0.159	0.743	0.397

effectively. Table 3 demonstrates the significant enhancement of the three student networks compared to baseline training. Notably, ESPNet and MobileNetV2 exhibit the most notable improvements, achieving performance increases of up to 6.7% (from 0.773 to 0.804) and 8.8% (from 0.714 to 0.802) respectively.

Our distillation pipeline enables ENet to surpass baseline performance in terms of segmentation accuracy, with ENet achieving a gain of 4.4% (from 0.808 to 0.848) in dice score. Across both experimental datasets, it is noteworthy that FCN with ResNet101 as the backbone emerges as the most outstanding teacher model, as all student networks experience the highest increase in learning compared to baseline training. Moreover, it is noteworthy that the Value Overlap Error (VOE) has decreased from the training without distillation to that with distillation for both organs, prostate, and spleen, as shown in Table 3.

The FLOPS and number of parameters, as depicted in Table 2, are computed utilizing the Flops counter in the PyTorch framework **ptflops**. This is achieved by supplying the input size, $1 \times 384 \times 384$, to the function **get_model_complexity_info** along with the trained model to interpret the computational complexity, as outlined in Table 2. Among all tested models, ESPNet exhibited the lowest FLOPs, with a value of $1.27G$. Distillation with a well-trained teacher notably enhanced the segmentation results while significantly reducing FLOPs and the number of parameters.

4.4.2. Multi-teacher distillation

We assess the effectiveness of our multi-teacher distillation framework with adaptive and ensemble learning by applying it to various combinations of teacher and student networks. We aim to determine whether students can accurately replicate the teachers' abilities and gain deeper insights by simultaneously distilling knowledge from multiple teachers. Our approach leverages the advantages of multi-resource data used in our experiments. Hence, we opt to train the student model by leveraging insights from three teacher networks for prostate data, while employing two teachers to guide the lightweight model trained on spleen data.

MRI Prostate segmentation: Considering the three sources providing MRI Prostate data, we selected the three most powerful teacher networks to train each one with a subset from the total data separately. Our selection of multi-teacher networks was guided by the teacher's ability to empower a lightweight student network to achieve the greatest enhancement. Consequently, we employed DeepLabV3+ [31], T2: FCN-ResNet101 [14], and T3: PSPNet [13] to implement the multi-teacher distillation architecture on Prostate data that yielded the highest dice scores in the mono-teacher KD. To ensure a thorough training methodology, we divided the overall dataset into three subsets, corresponding to the three distinct.

Table 4
Cross experiments results between multi-teachers and students on prostate and spleen data. T1, T2, and T3 refer to DeepLabV3+, FCN_ResNet101, and PSPNet respectively.

Teacher networks	Prostate		Spleen	
	Dice	VOE	Dice	VOE
T1 + subset1	0.899	0.119	0.922	0.057
T2 + subset2	0.912	0.078	0.957	0.041
T3 + subset3	0.874	0.186	–	–
Student Networks and their performances distilled from multi-teachers				
ENet: w/o	0.800	0.228	0.808	0.217
T1 + T2 + T3: w/	0.839	0.162	–	–
T1 + T2: w/	–	–	0.911	0.087
ESPNet: w/o	0.752	0.412	0.773	0.388
T1 + T2 + T3: w/	0.834	0.171	–	–
T1 + T2: w/	–	–	0.877	0.098
MobileNetV2: w/o	0.784	0.310	0.714	0.536
T1 + T2 + T3: w/	0.848	0.123	–	–
T1 + T2: w/	–	–	0.855	0.139

Furthermore, we implemented an adaptive weight determination method based on the dice loss for each teacher. This approach separately addresses both distillation levels: logits distillation loss and features distillation loss as described in Section 3.3. Table 4 presents the performance metrics of the student networks trained through distillation from multiple teachers. Remarkably, upon comparing with Table 3, it is evident that all three students ENet [3], ESPNet [2], and MobileNetV2 [44] demonstrate significant enhancements in terms of dice scores. This advancement is particularly noteworthy as the students effectively emulate the teaching capabilities of DeepLabV3+ [31], PSPNet [13], and FCN_ResNet101 [14], meticulously crafted for cutting-edge medical imaging segmentation. This meticulous methodology enabled us to assess and highlight the potential improvements achieved by the selected student networks in distillation scenarios involving multiple teachers across our experimental datasets. As shown in Table 4, all the student models demonstrated improved segmentation performance, achieving advancements of 3.9% (0.800 to 0.839), 8.2% (0.752 to 0.834), and 6.4% (0.784 to 0.848) for ENet, ESPNet, and MobileNetV2, respectively. Notably, ESPNet, despite being the most compact model, achieves a significantly higher dice score compared to ENet, with only a minor difference of 0.5% in dice score. Furthermore, in Table 4, it is noticeable that the Value Overlap Error (VOE) decreased from the baseline training to that after distillation. Moreover, it is evident that the multi-teacher distillation approach yields a dice score comparable to that of mono-teacher distillation. However, our objective extends beyond merely constructing a distillation pipeline that enables a lightweight model to replicate the performance of strong teachers from mono-teacher distillation. We aim to create a compact and adaptable model capable of handling multi-source data, addressing data privacy concerns, and being deployable on resource-constrained devices without incurring additional computational costs. Essentially, our objective is not to enable the student to achieve higher performance by combining multiple teachers compared to learning from a single teacher. It is crucial to emphasize that distilling from multiple teachers aims to enable the student to leverage the unique capabilities of each teacher independently. For Qualitative results, Fig. 5 visually presents segmentation masks of various samples. The first two rows depict the results of MobileNetV2 distilled from PSPNet and DeepLabV3+, respectively. Following that, the next two rows show the predicted

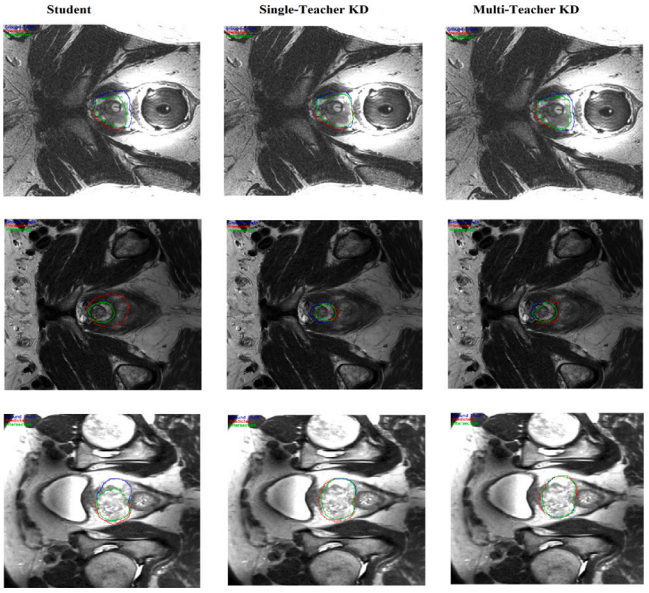


Fig. 4. Visualization of segmentation overlays for single and trio-teacher distillation on MRI prostate. The contours' colors blue, red, and green represent the GT, predicted mask, and the intersection, respectively.

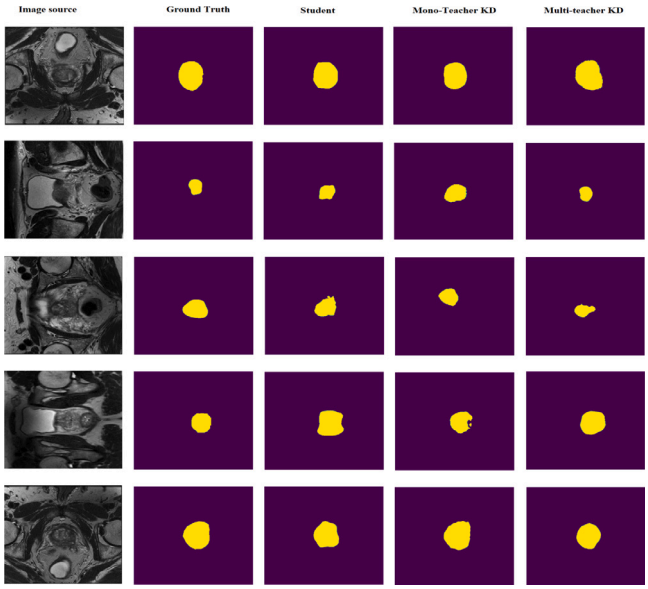


Fig. 5. Visualization of the prediction results through single and trio-teacher distillation on MRI prostate.

masks of ESPNet refined from PSPNet and FCN ResNet101 respectively. Lastly, the segmentation mask of ENet with distillation from FCN ResNet101 is visualized in the last row. Furthermore, there is a notable contrast among the segmentation masks produced by the student models in their baseline state (column 3) compared to when they were distilled from a single teacher model (column 4). In addition, Fig. 5 illustrates the segmentation results from multi-teacher distillation, shown in the last column, highlighting their enhanced discriminative and representative qualities compared to those from the baseline student and single-teacher distillation (columns 3 and 4 in Fig. 5 respectively). Interestingly, the third row presents an example where single-teacher distillation achieves superior segmentation performance compared to the multi-teacher method. Furthermore, in Fig. 4, we visualize the segmentation overlays of our predictions to highlight the performance of our distillation results. This finding underscores the efficacy of the multi-teacher distillation approach in medical imaging analysis, encompassing both segmentation and classification tasks. Despite the demonstrated improvements in student networks guided by teachers trained on limited datasets, the results emphasize the need for further enhancements in the multi-teaching scheme. Optimizing this approach could significantly improve prediction outcomes, representing a promising direction for our future research.

CT Spleen segmentation : Fascinatingly, utilizing a combination of CT Spleen data allows us to investigate dual teacher distillation, leveraging both Decathlon and Duke Data sources. For dual-teacher learning with Spleen data, we employed a similar approach as used for multi-teacher distillation with Prostate data. Referring to the results of single-teacher distillation depicted in Table 3, we pinpointed the two top-performing teachers, DeepLabV3+ and FCN ResNet101, based on their superior baseline dice scores. Subsequently, we chose these teachers to guide the student networks (ENet [3], ESPNet [2], and MobileNetV2 [44]), with the training process carried out on two separate subsets. Subset 1 comprised Duke data, while subset 2 encompassed Decathlon data. The quantitative results are displayed in Table 4, where MobileNetV2 demonstrates a significant improvement in dice score by 14.1% from the baseline (0.714) to the distillation (0.855). ESPNet and ENet also exhibit increases of up to 10.4% (0.773 to 0.877) and 10.3% (0.808 to 0.911), respectively. Compared to prostate segmentation, spleen segmentation with multi-teacher distillation has

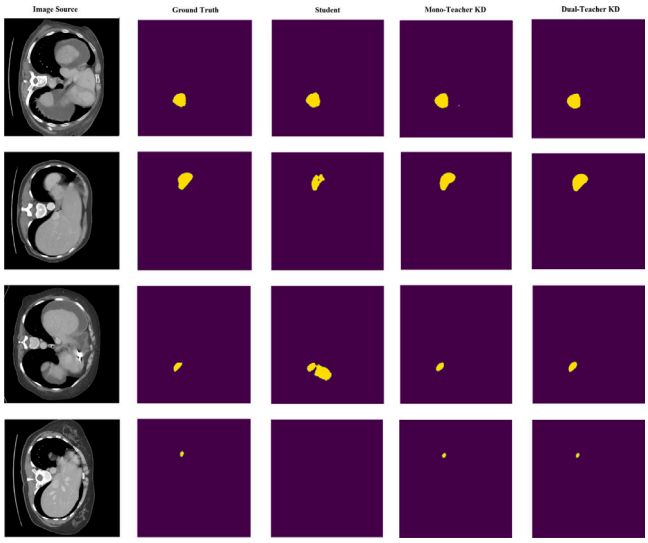


Fig. 6. Visualization of the prediction results through single and dual-teacher distillation on CT spleen.

improved tremendously due to the nature of spleen data. We collected this latter from two sites completely different in terms of institution, acquisition device, and thereby a huge distribution gap between the two sources. In Fig. 6, we present the visual forecasts of CT spleen in two scenarios: mono-teacher and dual-teacher. The first and third rows depict data from Duke, while the second and last rows showcase results from Decathlon. Notably, the advanced segmentation achieved through dual-teacher distillation is evident compared to the mono-teacher approach. This indicates a superior performance in delineating spleen tumors, particularly noticeable in the clarity and accuracy of the segmentation results. Additionally, by examining the segmentation overlays in Fig. 7, we can observe the intersection between the ground truth and our predictions. This clearly demonstrates the effectiveness of the dual-teacher distillation approach.

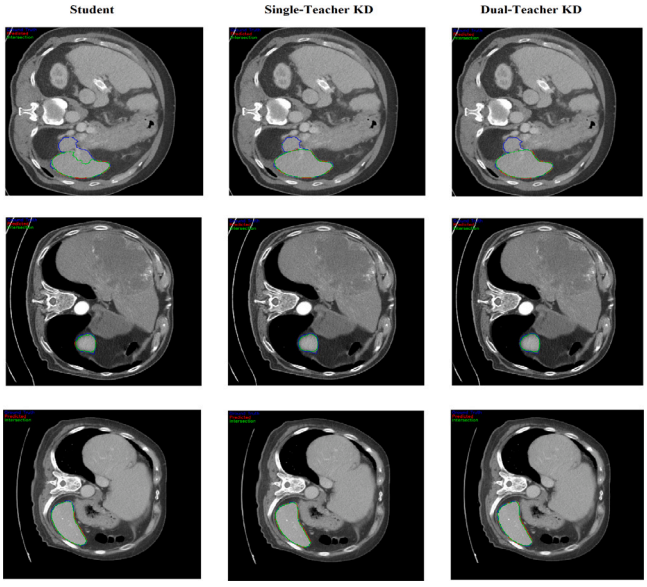


Fig. 7. Visualization of segmentation overlays for single and dual-teacher distillation on CT Spleen. The contours' colors blue, red, and green represent the GT, predicted mask, and the intersection respectively.

Table 5
Comparison with other knowledge distillation methods on Both prostate and spleen data. We select ENet and FCN_ResNet101 as the student and the teacher, respectively.

KD methods	Prostate	Spleen
	Dice score	Dice score
Student: ENet	0.800	0.808
Teacher: FCN_ResNet101	0.901	0.945
KD [4]	0.765	0.811
AT [7]	0.816	0.822
EMKD [7]	0.829	0.813
CIRKD [46]	0.826	0.824
ProtoKD [30]	0.770	0.846
AMTML-KD [10]	0.841	0.836
OURS-KD	0.859	0.848
OURS-MTKD	0.869	0.911

4.5. Comparisons with other KD methods

To highlight the effectiveness of our multi-teacher approach, it is necessary to evaluate it against several advanced distillation methods. These include widely used techniques like KD [4] and AT [34], as well as cutting-edge methods tailored for semantic segmentation, such as SKD [6], EMKD [7], CIRKD [46] and AMTML-KD [10]. To demonstrate our method's versatility, we conduct experiments using two distinct teacher networks. To illustrate the robustness of our method, we conducted experiments using ENet and FCN_ResNet101 as student and teacher networks respectively. This ensured that all distillation methods were evaluated under the same conditions. Table 5 showcases the comparison results on our experimental datasets, highlighting the clear benefits of our approach. Our dice scores outperform other methods in single-teacher and multi-teacher knowledge distillation scenarios. This indicates that our distillation technique allows the student network to effectively absorb detailed knowledge from one or more teachers, resulting in superior segmentation performance, particularly evident in spleen tumor segmentation with multi-teacher distillation.

5. Conclusion

To sum up, this work introduced two knowledge distillation approaches: single and multi-teacher schemes. The findings demonstrate the efficacy of our proposed multi-teacher distillation framework in addressing challenges related to medical data sharing and resource limitations, while simultaneously optimizing model performance with minimal computational costs. Our approach significantly outperformed state-of-the-art methods, achieving remarkable improvements in Dice scores by 9% for MRI prostate and 14.1% CT spleen segmentation in several instances. These results highlight the feasibility of preserving the accuracy of complex models trained on limited data within lightweight models suitable for deployment on resource-constrained devices. However, it is important to acknowledge certain limitations of our study. First, our research focused exclusively on 3D volumes, specifically MRI and CT scans, which may restrict the applicability of our findings to other medical imaging modalities. Future research should aim to extend our distillation framework to encompass a wider range of imaging techniques, such as ultrasound or X-ray, to enhance its versatility and impact. Moreover, we plan to explore the potential of multi-teaching online distillation, allowing for concurrent end-to-end training of both teacher and student networks within an online pipeline. This innovative approach is expected to improve the generalization of the student network and facilitate faster inference times for efficient and effective model deployment strategies in the medical imaging domain.

CRediT authorship contribution statement

Eddardaa Ben Loussaief: Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Hatem A. Rashwan:** Writing – review & editing, Supervision, Investigation. **Mohammed Ayad:** Writing – original draft, Methodology, Formal analysis. **Adnan Khalid:** Writing – original draft, Formal analysis. **Domemec Puig:** Validation, Supervision.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are not publicly available due to [privacy restrictions] but can be obtained from the corresponding author upon reasonable request and with appropriate institutional approvals.

References

[1] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.
[2] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, H. Hajishirzi, ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 561–580.
[3] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, ENet: A deep neural network architecture for real-time semantic segmentation, 2016.
[4] G. Hinton, J. Dean, O. Vinyals, Distilling the knowledge in a neural network, 2014, pp. 1–9.
[5] H. Tong, C. Shen, Z. Tian, D. Gong, C. Sun, Y. Yan, Knowledge adaptation for efficient semantic segmentation, 2019, pp. 578–587, http://dx.doi.org/10.1109/CVPR.2019.00067.
[6] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 2599–2608, http://dx.doi.org/10.1109/CVPR.2019.00271.
[7] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, H.-F. Dai, Efficient medical image segmentation based on knowledge distillation, IEEE Trans. Med. Imaging 40 (12) (2021) 3820–3831, http://dx.doi.org/10.1109/TMI.2021.3098703.

[8] L. Zhao, X. Qian, Y. Guo, J. Song, J. Hou, J. Gong, MSKD: Structured knowledge distillation for efficient medical image segmentation, Comput. Biol. Med. 164 (2023) 107284, http://dx.doi.org/10.1016/j.compbiomed.2023.107284.

[9] Y. Wen, L. Chen, S. Xi, Y. Deng, X. Tang, C. Zhou, Towards efficient medical image segmentation via boundary-guided knowledge distillation, 2021, pp. 1–6, http://dx.doi.org/10.1109/ICME51207.2021.9428395.

[10] Y. Liu, W. Zhang, J. Wang, Adaptive multi-teacher multi-level knowledge distillation, Neurocomputing 415 (2020) 106–113, http://dx.doi.org/10.1016/j.neucom.2020.07.048.

[11] H. Zhang, D. Chen, C. Wang, Adaptive multi-teacher knowledge distillation with meta-learning, in: 2023 IEEE International Conference on Multimedia and Expo, ICME, 2023, pp. 1943–1948, URL https://api.semanticscholar.org/CorpusID:259138547.

[12] Z. Huang, Z. Wang, J. Chen, Z. Zhu, J. Li, Real-time colonoscopy image segmentation based on ensemble knowledge distillation, in: 2020 5th International Conference on Advanced Robotics and Mechatronics, ICARM, 2020, pp. 454–459, http://dx.doi.org/10.1109/ICARM49381.2020.9195281.

[13] Z. Tian, L. Liu, Z. Zhang, B. Fei, PSNet: prostate segmentation on MRI based on a convolutional neural network, J. Med. Imaging 5 (2) (2018) 021208, http://dx.doi.org/10.1117/1.JMI.5.2.021208.

[14] Y. Liu, K. Sung, G. Yang, S. Afshari Mirak, M. Hosseiny, A. Azadikhah, X. Zhong, R. Reiter, Y. Lee, S. Raman, Automatic prostate zonal segmentation using fully convolutional network with feature pyramid attention, IEEE Access PP (2019) 1, http://dx.doi.org/10.1109/ACCESS.2019.2952534.

[15] Q. Zhu, B. Du, P. Yan, Boundary-weighted domain adaptive neural network for prostate MR image segmentation, IEEE Trans. Med. Imaging 39 (2019) 753–763.

[16] Q. Liu, Q. Dou, L. Yu, P. Heng, MS-net: Multi-site network for improving prostate segmentation with heterogeneous MRI data, IEEE Trans. Med. Imaging PP (2020) 1, http://dx.doi.org/10.1109/TMI.2020.2974574.

[17] R. Li, X. Chen, An efficient interactive multi-label segmentation tool for 2D and 3D medical images using fully connected conditional random field, Comput. Methods Programs Biomed. 213 (2021) 106534, http://dx.doi.org/10.1016/j.cmpb.2021.106534.

[18] H. Moon, Y. Huo, R. Abramson, R. Peters, A. Assad, T. Moyo, M. Savona, B. Landman, Corrigendum to "acceleration of spleen segmentation with end-to-end deep learning method and automated pipeline" [Comput. Biol. Med. 107 (2019) 109–117], Comput. Biol. Med. 140 (2022) 103684, http://dx.doi.org/10.1016/j.compbiomed.2020.103684.

[19] A. Mihaylova, V. Georgieva, Spleen segmentation in MRI sequence images using template matching and active contours, Procedia Comput. Sci. 131 (C) (2018) 15–22, http://dx.doi.org/10.1016/j.procs.2018.04.180.

[20] Y. Huo, Z. Xu, S. Bao, C. Bermudez, A.J. Plassard, J. Liu, Y. Yao, A. Assad, R.G. Abramson, B.A. Landman, Splenomegaly segmentation using global convolutional kernels and conditional generative adversarial networks, in: E.D. Angelini, B.A. Landman (Eds.), in: Medical Imaging 2018: Image Processing, vol. 10574, SPIE, International Society for Optics and Photonics, 2018, 1057409, http://dx.doi.org/10.1117/12.2293406.

[21] A. Comelli, N. Dahiya, A. Stefano, F. Vernuccio, M. Portoghese, G. Cutaia, A. Bruno, G. Salvaggio, A. Yezzi, Deep learning-based methods for prostate segmentation in magnetic resonance imaging, Appl. Sci. 11 (2021).

[22] L. Ma, X. Hou, Z. Gong, Image segmentation technology based on attention mechanism and ENet, Comput. Intell. Neurosci. 2022 (2022) 1–8, http://dx.doi.org/10.1155/2022/9873777.

[23] N. Nuechterlein, S. Mehta, 3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation, in: A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, T. van Walsum (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, 2019.

[24] Y. Wang, Y. Wang, J. Cai, T. Lee, C. Miao, Z. Wang, SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images, Med. Image Anal. 84 (2022) 102693, http://dx.doi.org/10.1016/j.media.2022.102693.

[25] Q. Dou, Q. Liu, P. Heng, B. Glocker, Unpaired multi-modal segmentation via knowledge distillation, IEEE Trans. Med. Imaging PP (2020) 1, http://dx.doi.org/10.1109/TMI.2019.2963882.

[26] K. Li, L. Yu, S. Wang, P.-A. Heng, Towards cross-modality medical image segmentation with online mutual knowledge distillation, Proc. AAAI Conf. Artif. Intell. 34 (2020) 775–783, http://dx.doi.org/10.1609/aaai.v34i01.5421.

[27] Y. Qi, W. Zhang, X. Wang, X. You, S. Hu, J. Chen, Efficient knowledge distillation for brain tumor segmentation, Appl. Sci. 12 (23) (2022) http://dx.doi.org/10.3390/app122311980, URL https://www.mdpi.com/2076-3417/12/23/11980.

[28] J.M.H. Noothout, N. Lessmann, M.C.V. Eede, L.D. van Harten, E. Sogancioglu, F.G. Heslinga, M. Veta, B. van Ginneken, I. Išgum, Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation, J. Med. Imaging 9 (5) (2022) 052407, http://dx.doi.org/10.1117/1.JMI.9.5.052407.

[29] J.M.H. Noothout, N. Lessmann, M.C.V. Eede, L.D. van Harten, E. Sogancioglu, F.G. Heslinga, M. Veta, B. van Ginneken, I. Išgum, Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation, J. Med. Imaging 9 (5) (2022) 052407, http://dx.doi.org/10.1117/1.JMI.9.5.052407.

[30] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, R. Li, Prototype knowledge distillation for medical segmentation with missing modality, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2023, http://dx.doi.org/10.1109/ICASSP49357.2023.10095014.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. PP (2016) http://dx.doi.org/10.1109/TPAMI.2017.2699184.

[32] A.L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. van Ginneken, A. Kopp-Schneider, B.A. Landman, G. Litjens, B. Menze, O. Ronneberger, R.M. Summers, P. Bilic, P.F. Christ, R.K.G. Do, M. Gollub, J. Golia-Pernicka, S.H. Heckers, W.R. Jarnagin, M.K. McHugo, S. Napel, E. Vorontsov, L. Maier-Hein, M.J. Cardoso, A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019, arXiv:1902.09063.

[33] Y. Wang, J.A. Macdonald, K.R. Morgan, D. Hom, S. Cubberley, K. Sollace, N. Casasanto, I.H. Zaki, K.J. Lafata, M.R. Bashir, Duke spleen data set: A publicly available spleen MRI and CT dataset for training segmentation, 2023, arXiv:2305.05732.

[34] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2017, arXiv:1612.03928.

[35] M. Berman, A. Rannen, M. Blaschko, The Lovasz-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks, 2018, pp. 4413–4421, http://dx.doi.org/10.1109/CVPR.2018.00464.

[36] S. You, C. Xu, C. Xu, D. Tao, Learning from Multiple Teacher Networks, Association for Computing Machinery, New York, NY, USA, 2017, http://dx.doi.org/10.1145/3097983.3098135.

[37] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, B. Ramabhadran, Efficient knowledge distillation from an ensemble of teachers, in: Interspeech 2017, 2017, pp. 3697–3701, http://dx.doi.org/10.21437/Interspeech.2017-614.

[38] X. Tan, Y. Ren, D. He, T. Qin, Z. Zhao, T.-Y. Liu, Multilingual neural machine translation with knowledge distillation, 2019, arXiv:1902.10461.

[39] J. Vongkulbhisal, P. Vinayavekhin, M. Visentini-Scarzanella, Unifying heterogeneous classifiers with distillation, 2019, arXiv:1904.06062.

[40] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, A. Madabhushi, Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge, Med. Image Anal. 18 (2013) 359–373, http://dx.doi.org/10.1016/j.media.2013.12.002.

[41] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, K. Farahani, NCI-ISBI 2013 challenge: Automated segmentation of prostate structures, 2015, http://dx.doi.org/10.7937/K9/TCIA.2015.zF0vlOPv, the Cancer Imaging Archive.

[42] G. Lemaître, R. Martí, J. Freixenet, J.C. Vilanova, P. Walker, F. Meriaudeau, Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review, Comput. Biol. Med. (2015) http://dx.doi.org/10.1016/j.compbiomed.2015.02.009.

[43] S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, 2017, pp. 5987–5995, http://dx.doi.org/10.1109/CVPR.2017.634.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, 2018, pp. 4510–4520, http://dx.doi.org/10.1109/CVPR.2018.00474.

[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[46] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, Q. Zhang, Cross-image relational knowledge distillation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12319–12328.