scientific reports

Check for updates

OPEN

Enhancing efficient deep learning models with multimodal, multi-teacher insights for medical image segmentation

Khondker Fariha Hossain[1]✉, Sharif Amit Kamran[1], Joshua Ong[2] & Alireza Tavakkoli[1]

The rapid evolution of deep learning has dramatically enhanced the field of medical image segmentation, leading to the development of models with unprecedented accuracy in analyzing complex medical images. Deep learning-based segmentation holds significant promise for advancing clinical care and enhancing the precision of medical interventions. However, these models' high computational demand and complexity present significant barriers to their application in resource-constrained clinical settings. To address this challenge, we introduce Teach-Former, a novel knowledge distillation (KD) framework that leverages a Transformer backbone to effectively condense the knowledge of multiple teacher models into a single, streamlined student model. Moreover, it excels in the contextual and spatial interpretation of relationships across multimodal images for more accurate and precise segmentation. Teach-Former stands out by harnessing multimodal inputs (CT, PET, MRI) and distilling the final predictions and the intermediate attention maps, ensuring a richer spatial and contextual knowledge transfer. Through this technique, the student model inherits the capacity for fine segmentation while operating with a significantly reduced parameter set and computational footprint. Additionally, introducing a novel training strategy optimizes knowledge transfer, ensuring the student model captures the intricate mapping of features essential for high-fidelity segmentation. The efficacy of Teach-Former has been effectively tested on two extensive multimodal datasets, HECKTOR21 and PI-CAI22, encompassing various image types. The results demonstrate that our KD strategy reduces the model complexity and surpasses existing state-of-the-art methods to achieve superior performance. The findings of this study indicate that the proposed methodology could facilitate efficient segmentation of complex multimodal medical images, supporting clinicians in achieving more precise diagnoses and comprehensive monitoring of pathological conditions (https://github.com/FarihaHossain/TeachFormer).

Medical image segmentation is crucial for precise clinical diagnoses and cancer progression assessment but encounters challenges due to the appearance of the organ and tumor diversity, irregular sizes, and unpredictable locations. Deep learning, particularly the U-Net[1] architecture, has significantly advanced medical image segmentation by enabling more accurate segmentation of precise regions and tumor tissues across various imaging modalities. In these advancements, Unet-based architectures like RA-UNet[2], CEU-Net, and transformer-based architectures like Trans-UNet[3], UNetr[4], Swin-Unet[5] have been implemented to enhance the target regions and the representation of features. Yet, despite this technological progress, the majority of cutting-edge semantic segmentation models require extensive computational power, restricting their feasible use in clinical settings. Moreover, Collecting extensive and well-curated data with ground truth annotation is challenging in underprivileged areas despite the abundance of available data[6,7]. Furthermore, medical image segmentation is crucial for precise clinical diagnoses and cancer progression assessment but encounters challenges due to the appearance of the organ and tumor diversity, irregular sizes, and unpredictable locations[8–11].

Deploying segmentation models in real-world scenarios is challenging due to their substantial computational demands. Lightweight networks like 3D UX-Net[12], 3D Medical Axial Transformer[13] are aimed at real-time semantic segmentation and attention toward real-time medical image segmentation. However, Xiaogang Du et al.[14] pointed out the underlying consequence that these simplified models can often compromise

[1]Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV 89557, USA. [2]Department of Ophthalmology and Visual Sciences, University of Michigan Kellogg Eye Center, Ann Arbor, MI, USA. ✉email: khondkerfarihah@gmail.com

their performance, emphasizing in the medical image that accurate segmentation is crucial in this domain. Techniques such as model compression[15], transfer learning[16], and knowledge distillation[17] have been explored to address these limitations. Knowledge distillation, in particular, has gained attention for transferring information from well-trained teacher networks to lightweight student networks to enhance performance without the computational overhead[6,18,19] (Fig. 1).

Meanwhile, transformer models rely on self-attention mechanisms to capture long-range dependencies and complex patterns in data, resulting in a large number of parameters. Though beneficial for capturing complex patterns, it can result in overfitting, especially in scenarios with limited training data, which is a common issue[19]. In corporating, Transformers with KD strategy can offer a state-of-the-art platform for distilling complex, high-dimensional knowledge from teacher models to more compact student models. This approach facilitates a deeper understanding of spatial and contextual relationships within multimodal medical images to ensure that the distilled student models retain precision and accuracy, hence a lightweight model[20–22]. However, conventional distillation methods need to pay more attention to the rich information available during the learning process and need help maintaining fine-grained semantic information. While these techniques offer viable solutions to alleviate the initial challenges, they frequently result in more complex and burdensome architecture. Considering all the relevant factors, we propose the following:

- We present an innovative approach, Teach-Former, to knowledge distillation specifically designed for multi-modal medical image segmentation. This method uniquely features both the teacher and student networks and the ability to dynamically learn from teachers towards the shared knowledge during the training process. Our approach incorporates knowledge distillation from multiple teachers to segment outputs with multi-modal imaging data, including Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI).
- The training strategy distills knowledge from intermediate feature-level attention maps and output prediction maps from multiple teachers to a single student model. While previous approaches suffered from scaling intermediate features from different parts of the encoder or decoder to calculate the intermediate feature loss, our coarse and fine attention map distillation loss utilizes feature extraction with uniform dimensions from the same encoder and decoder layers (Fig. 2).
- Furthermore, we provide extensive results that showcase superior performance over existing state-of-the-art approaches in terms of parameter reduction and higher dice score gain (Fig. 1).

Key contribution

1. Leverages diverse spatial and contextual knowledge from multiple teachers using multi-modal inputs (CT, PET, MRI).
2. Introduces coarse and fine attention similarity loss for effective intermediate knowledge transfer.
3. Achieves 5× to 10× parameter reduction and 10× to 15× lower GFLOPs while maintaining high accuracy.
4. Demonstrates superior performance on HECKTOR’21 and PI-CAI’22 datasets compared to state-of-the-art methods.

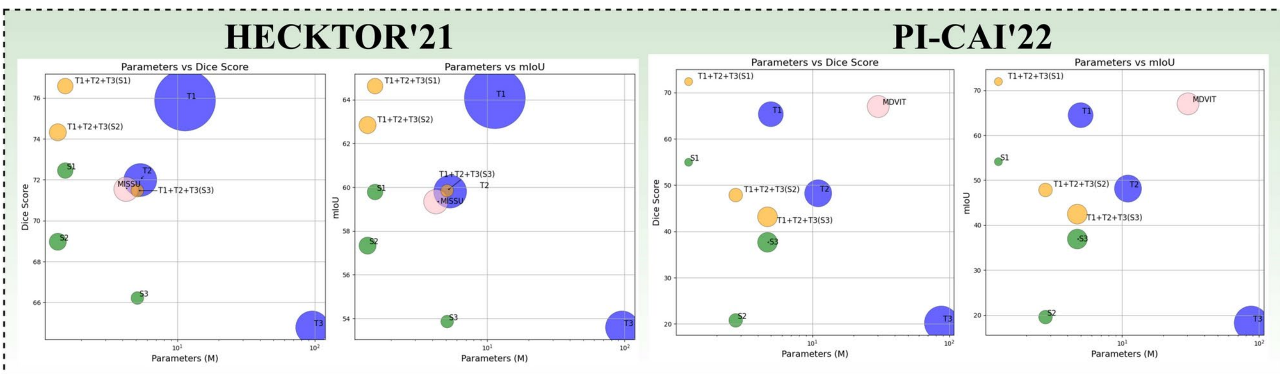

Fig. 1. The figure compares the performance of various models in terms of Dice Score and mIoU against the number of parameters (in millions) across two datasets: HECKTOR’21 (left) and PI-CAI’22 (right). The bubble size represents the computational complexity of each model, measured in GFLOPs. Models are categorized into four groups: Teacher models (T1, T2, T3) shown in blue, Student models (S1, S2, S3) in green, Other KD approaches (MISSU, MDVIT) in pink, and Aggregated models derived from multiple teachers (T1+T2+T3(S1), T1+T2+T3(S2), T1+T2+T3(S3)) in orange. The plots demonstrate how different models balance accuracy and computational efficiency, with the aggregated models generally achieving higher performance with fewer parameters.

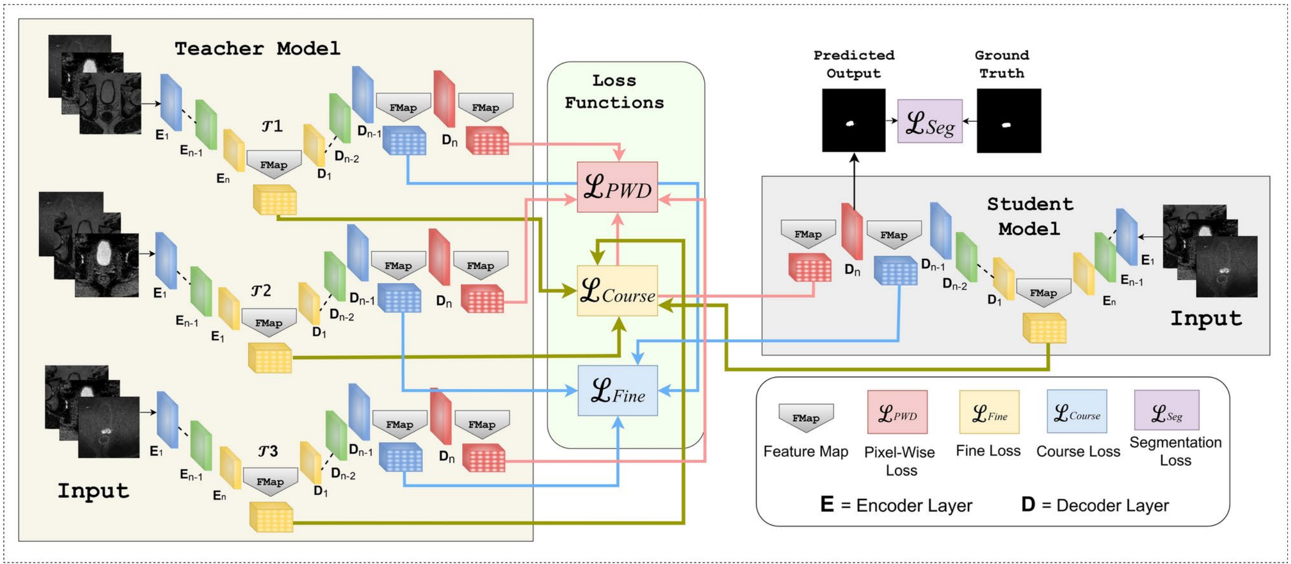

Fig. 2. Our proposed architecture, Teach-Former, consisting of teacher models *T*1, *T*2 and *T*3 and one student model. The teacher models are utilized in conjunction with the student model to calculate coarse and fine attention similarity and pixel-wise distillation loss. Furthermore, the student model is trained using a segmentation loss consisting of focal and dice loss.

Related work
Knowledge distillation in medical imaging
Knowledge distillation (KD) has emerged as a pivotal technique for enhancing the performance of lightweight models by transferring knowledge from larger, more complex teacher models. Hinton et al.'s pioneering work on KD demonstrated its potential to maintain high accuracy in student models while significantly reducing computational demands. Various studies have leveraged KD to balance the trade-off between model complexity and performance in medical image segmentation. Knowledge distillation (KD) has emerged as an essential technique for developing lightweight models, crucial in medical fields where computational resources are limited. This approach transfers knowledge from a complex, resource-intensive teacher model to a more efficient student model, maintaining high performance while reducing computational demands. In clinical settings, lightweight models are vital for real-time processing, efficient use of resources, and reliable performance. For instance, Zhong et al.[23] highlight using deep learning models for auto segmentation in head and neck cancer, emphasizing the need for high accuracy and efficiency in typical clinical hardware. Similarly, Kosmin et al.[24] discuss advancements in the auto-segmentation of organs at risk and target volumes in radiotherapy, where balancing performance with computational efficiency is crucial for integration into clinical workflows. Kitaguchi et al.[25] demonstrate the feasibility of real-time automatic prostate segmentation during TaTME, underscoring the necessity for accurate real-time segmentation in surgical settings. Again, RSKD[26] improves segmentation with rank-sensitive distillation but prioritizes fine-grained features without effectively addressing computational efficiency or handling multi-modal data. Various KD strategies, such as attention transfer, multi-teacher approaches, and intermediate layer distillation, enhance the student model's ability to perform complex tasks efficiently, addressing the practical challenges in medical image segmentation.

Multimodal medical imaging
Integrating information from different imaging modalities, such as CT, PET, and MRI, presents unique challenges and opportunities in medical imaging. Multimodal approaches leverage the complementary information different imaging techniques provide to enhance segmentation performance. Valindria et al.[27] demonstrated that combining CT and MRI data significantly improves multi-organ structure segmentation accuracy. Similarly, Zhou et al.[28] proposed the Unet++ architecture, which benefits from multimodal data integration to achieve more precise medical image segmentation. Martí-Bonmatí et al. emphasize that each imaging modality-including CT, PET, MRI, and others-offers distinct information that can be vital for precise diagnosis and treatment strategizing. For instance, CT scans are excellent for visualizing bone structures, while MRI is superior for soft tissue contrast. PET scans, on the other hand, provide metabolic information vital for identifying cancerous tissues. However, integrating these different data types presents significant technical challenges, including the need for sophisticated image registration, data fusion techniques, and the handling of large volumes of data[29]. Guo et al. review the application of multimodal deep learning in medical imaging, underscoring its potential to enhance visual understanding and improve clinical outcomes. The discussion revolves around training deep learning models to combine data from various imaging modalities, enabling them to capture a more comprehensive range of features that single-modality models could overlook. This multimodal strategy can potentially enhance the accuracy of diagnostic tools and predictive models, thereby increasing their reliability for clinical applications. Moreover, Guo et al.[30] highlight the advancements in neural

network architectures that facilitate multimodal learning, such as convolutional neural networks (CNNs) and transformers, which can handle complex, high-dimensional data. Despite these advancements, they also highlight ongoing challenges, such as the need for large, annotated datasets and the computational complexity involved in training and deploying these models[30]. Li et al.'s[31] work on knowledge distillation addresses the need for efficient models that maintain high performance in resource-constrained clinical settings. Similarly, Song et al.[32] combine residual U-Nets and Vision Transformers in a teacher-student framework, focusing more on classification tasks and less on transferring spatial and contextual knowledge. While these methods contribute to specific areas, they generally do not integrate multi-teacher frameworks or optimize intermediate feature-level distillation for diverse multi-modal datasets, presenting a gap addressed by our proposed approach. These studies underscore the critical importance of multimodal imaging in improving diagnostic accuracy and treatment planning, emphasizing the necessity for advanced techniques to effectively distill and integrate this information into streamlined, efficient models for clinical application.

Attention features with 2D and 3D image modality

With their self-attention mechanisms, transformer models have revolutionized various areas of machine learning by effectively capturing long-range dependencies and intricate patterns. Architectures like Trans-UNet[3], UNETR[4], and Swin-Unet[5], based on transformers, have shown great promise in medical image segmentation. These architectures are adept at capturing complex details and dependencies within medical images, making them highly suitable for segmentation tasks. However, their considerable computational requirements and model complexity pose significant challenges for practical implementation in clinical settings. The high computational load and susceptibility to overfitting on limited training data hinder their use in environments with constrained computational resources. Integrating transformers with knowledge distillation (KD) strategies offers a novel approach to leverage their powerful representation capabilities while mitigating overfitting through effective knowledge transfer, potentially making these advanced models more feasible for clinical use[3,18].

RA-UNet[2] utilizes attention mechanisms to enhance feature representation, offering improved segmentation accuracy, but it is restricted to single-modal data, limiting its generalizability. Similarly, Trans-UNet[3] employs Transformer-based encoders to effectively capture global dependencies; however, its high computational demands and lack of scalability for multi-modal inputs pose challenges in resource-constrained settings. Both architectures, while innovative, struggle to fully address the complexities of multi-modal medical image segmentation and the need for lightweight, efficient models adaptable to diverse clinical environments. KD can offer significant benefits in the realm of 2D and 3D medical image data. In recent work, the author proposes MDViT[33], which utilizes a multi-domain network trained on four different skin lesion image datasets using auxiliary peers through mutual knowledge distillation. The author incorporated domain-specific labels (0-4) as input alongside the images during model training. This approach addresses data scarcity and mitigates negative knowledge transfer (NKT) in scenarios involving multiple small datasets (domains). Although the model demonstrated good performance on 2D skin lesion images with limited datasets, it has notable limitations. Specifically, the model requires domain-specific labels during both training and inference, making it incapable of independently predicting the domain. Moreover, its effectiveness on 3D MRI or CT scan volumes was not evaluated, leaving its performance on higher-dimensional datasets unexplored. Furthermore, for single-domain training, the model's auxiliary peers has no utility, which becomes redundant in such scenarios. In another work, Wang et al.[34] proposed MISSU, which leverages Vision Transformers and self-distillation for medical image segmentation. The authors introduced a multi-scale fusion (MSF) block, integrated hierarchically atop the encoder. Feature-wise distillation loss was computed between the features from the encoder and the MSF blocks. The encoder features were added to the decoder features via skip connections, while the MSF block features were excluded from the decoder during both training and inference. A significant limitation of MISSU is that it was evaluated solely on MRI modalities, with no experiments conducted on PET or CT modalities for multi-modal training. Additionally, the MSF block is not utilized during testing or inference, which may reduce its practical utility. For 2D medical images from MRI or CT scans, KD aids in reducing model size and complexity, enabling faster inference times and smoother integration into clinical workflows[1,35]. The computational demands are even greater for 3D medical images, which involve more complex data structures and larger volumes. KD can efficiently distill essential features and spatial relationships from high-dimensional 3D data into more compact models without substantial loss of accuracy[36,37]. This capability ensures that advanced segmentation models can be utilized effectively in diverse clinical scenarios, showcasing their adaptability and versatility in the field of medical imaging.

Methods

Overall architecture

In this section, we describe our proposed architecture, Teach-Former, illustrated in Fig. 2 detail. Our proposed knowledge distillation approach consists of three transformer-based teachers and one student model. The architectures take multi-modal input CT and PET scans for the Hecktor21 dataset and three types of MR scans for the PI-CAI22 dataset. For teacher models, we utilize three recent architectures which reached state-of-the-art performance in spatial and volumetric medical image segmentation tasks, namely H-Denseformer[38], SwinUNETR[39], and UNETR[4]. In contrast, we develop three light-weight student models based on H-Denseformer, SwinUNETR, and 3DUX-Net[12]. The reason for utilizing 3DUX-Net is that it is similar in architecture to UNETR despite having $19\times$ less parameters. All the teacher models are pre-trained on the same dataset, so we freeze all the weights and parameters for distillation training. To distill strong spatial and depth feature information, we utilize coarse and fine attention maps from teachers to calculate similarities with our student model's coarse and fine attention maps. The coarse attention maps are extracted from the E_n encoder layer of the teacher and student models. At the same time, the fine features are extracted from the D_{n-1} decoder layer of the teacher and student models. We utilize these features to calculate the coarse and fine attention similarity distillation, which is described in section "Attention Feature Similarity Distillation". Next, we extract the pixel-wise output map from the D_n decoder layer of the teacher and student models and calculate the pixel-wise distillation loss. This loss helps the student architecture to emulate the output of the final layer of the teacher to learn segmentation prediction faster and more accurately and is elaborated in section "Pixel-wise Distillation". Finally, we utilize the combined focal[40] and dice loss function corresponding to the segmentation task, which ensures a primary convergence task corresponding to the multi-modal inputs and is described in section "Segmentation Loss Function".

Attention feature similarity distillation

Distilling knowledge from the intermediate layers is equally important as learning from the final layer of the teacher of the model[6,18]. The primary impediment of auto-encoder-based architecture is that the depth or spatial dimension of the teacher model's layer-wise attention feature does not match the corresponding feature dimension of the student model. Previously, various techniques have been employed, such as using different layers for feature extraction and resizing the features in the teacher and student models[6,18,41] for distilling knowledge. As a result, it contributes to inconsistent attention maps for loss convergence due to interpolating across spatial, depth, and feature channel dimensions. In this study, we address this limitation by introducing uniform down-sampling and up-sampling strategies across three teacher models and one student model, creating equal sizes for spatial (2D images) and depth dimensions (3D volumes). Moreover, we only interpolate the feature channel dimension because the student model has fewer feature channels than teachers. For 3D volumes, We take coarse attention map, $P \in \mathbb{R}^{K \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$ from layer, E_n, and fine attention map, $Q \in \mathbb{R}^{C \times H \times W \times D}$ from layer D_{n-1} of the teacher and student models. Here, *K* and *C* channel values depend on the teacher and their corresponding student network's encoder and decoder layers. By combining all the coarse and fine attention maps, we can calculate the proposed attention feature similarity distillation as follows:

$$\mathscr{L}_{AFS} = \sum_{t=1}^{T} \| \frac{P_s}{\|P_s\|_1} - \frac{P_t}{\|P_t\|_1} \|_2 + \| \frac{Q_s}{\|Q_s\|_1} - \frac{Q_t}{\|Q_t\|_1} \|_2 \tag{1}$$

In Eq. (1), we use $T = 3$ for the three teachers, $\| \cdot \|_1$ is *l*1 norm and $\| \cdot \|_2$ is *l*2 norm.

Pixel-wise distillation

The fundamental method of knowledge distillation[17] tries to drive the student network to acquire knowledge from the teacher network by calculating the difference of their output probability map with either cross-entropy or Kullback–Leibler divergence loss function. Motivated by this distillation approach, we incorporate a similar knowledge distillation strategy for semantic segmentation[42,43] to construct the Pixel-wise Distillation loss. Given the final output layer, D_n of the teacher and student models outputs a pixel-wise map, $O \in \mathbb{R}^{H \times W \times D}$, the corresponding loss function can be calculated as follows:

$$\mathscr{L}_{PWD} = \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i \in N} KL(O_i^s \| O_i^t) \right) \tag{2}$$

Here, in Eq. (2), $N =$ number of pixels, *KL*= Kullback-Leibler divergence loss, $T = 3$, is the number of teachers. Moreover, O^s and O^t represent the probabilities of the *i*th pixel in the segmentation map extracted from the student and the multiple teachers network successively.

Segmentation loss function

For binary segmentation tasks, dice or meanIOU loss is generally utilized. However, due to the dataset having more background pixels compared to the foreground tumor pixels, we need to address the pixel imbalance problem.To address the challenges inherent in medical image segmentation, we employ a combined focal and dice loss function. Focal loss is particularly effective in handling the significant class imbalance commonly observed in medical datasets, where background pixels often far outnumber the target pixels like tumor or organ regions. By dynamically down-weighting the loss contribution of well-classified examples, focal loss ensures greater emphasis on harder-to-classify pixels. On the other hand, dice loss directly optimizes for the overlap between the predicted and ground truth segmentation masks, making it especially suited for capturing fine-grained boundary details. For this, we use the joint Focal[40] and Dice loss as the convergence criterion for the segmentation task and is given as follows :

$$\mathscr{L}_{SEG} = 1 - \frac{2 \sum_{i=1}^{N} p_i t_i}{\sum_{i=1}^{N} p_i + t_i} + \frac{1}{N} \sum_{i=1}^{N} -(1-p_i)^{\gamma} log(p_i) \tag{3}$$

In Eq. (3), $N =$ the number of pixels, p_i and t_i is the *i*th predicted and ground truth pixels. By combining all Eqs. (1)–(3), we can devise our final objective function as follows:

$$\mathscr{L} = \mathscr{L}_{AFS} + \mathscr{L}_{PWD} + \mathscr{L}_{SEG} \tag{4}$$

We do not use any weighting mechanism to prioritize any particular loss. Moreover, we use early stopping with patience $= 40$ to stop training if the validation loss does not improve for forty consecutive epochs.

Results

Dataset

While training our models, we focused on datasets encompassing a broad range of imaging techniques, not limited to various MR sequences but including others like CT and PET. We also considered that the datasets should facilitate evaluation in 2D and 3D versions. We conducted experiments on the HECKTOR21[44] and PI-CAI22[45] datasets. The HECKTOR21 dataset is for head and neck tumor segmentation, consisting of 224 PET-CT image pairs aligned and trimmed to a uniform size ($144 \times 144 \times 144$). Again, the PI-CAI22 has 220 patient MRI images of prostate cancer, including T2-weighted (T2W), Diffusion-Weighted Imaging (DWI) at high b-value, and maps of Apparent Diffusion Coefficient (ADC) with the uniform dimensions of ($24 \times 384 \times 384$). The PI-CAI22 dataset, in particular, was favored for its alignment with 2D assessments due to its reduced slice count per sample, so the volumes were divided into 24 scans (each having 384×384 resolution). We selected 180 samples from all images by resampling and central cropping and trained in the process of five-fold cross-validation. In contrast, the remaining samples were reserved for independent test-set evaluation (44 from HECKTOR21 and 40 from PI-CAI22).

Hyper-parameter

We utilized the Adam optimizer for training due to its adaptive learning rate mechanism and proven effectiveness in medical image segmentation tasks, as highlighted in prior studies[46–48].The learning rate was set to $\sigma = 1e^{-3}$, a commonly validated value, ensuring stable convergence and optimal performance across various segmentation models. All networks in Teach-Former are trained using weight decay of $wd = 1e^{-4}$. To minimize the risk of overfitting, we incorporated online data augmentation practices, including various random rotations and flips, throughout the training phase. Furthermore, we trained all the models for 100 epochs, and to capture the best results, we used the strategy of Early Stopping with a patience of 40 epochs. We utilized the Polynomial Learning Rate method for the optimizer to manage the learning rate throughout the training process. For training and evaluation, all the Teachers' and Students' models were trained from scratch in NVIDIA A-30 GPUs with batch sizes of $b = 2$ for 3D images and $b = 16$ for 2D images using Pytorch.

Quantitative analysis

We incorporate our Teach-Former knowledge distillation approach with some best-performing 3D and 2D Transformer architectures for segmentation, including H-DenseFormer[38], MISSU[34], MDVIT[33], 3DUXNet[12], UNETR[4], and SwinUNETR[39], as illustrated in Table 1. We trained and evaluated these architectures using their

HECKTOR (3D)				
Model	**Param(M) ↓**	**GFLOPs ↓**	**Dice ↑**	**mIoU ↑**
T1: H-DenseFormer	11.38	960.38	75.88	64.07
T2: Swin-UNETR	5.38	280.91	71.99	59.80
T3: UNETR	95.76	282.18	64.78	53.59
MISSU	4.24	154.38	71.54	59.34
S1: H-Dense Former (Light)	1.53	62.73	72.45	59.78
S2: Swin-UNETR (Light)	1.35	75.68	68.97	57.33
S3: 3DUXNet	5.12	41.51	66.22	53.86
Aggregated Student performance derived from multiple Teachers (T1, T2, T3)				
H-Dense Former (Light)	1.53	62.73	76.58	64.62
Swin-UNETR (Light)	1.35	75.68	74.32	62.84
3DUXNet	5.12	41.51	71.47	59.84
PI-CAI (2D)				
T1: H-DenseFormer	4.98	32.33	65.33	64.46
T2: Swin-UNETR	11.04	37.94	48.23	48.13
T3: UNETR	87.51	59.42	20.21	18.24
MDVIT	30.30	24.95	66.99	66.98
S1: H-Dense Former (Light)	1.25	3.31	54.94	54.11
S2: Swin-UNETR (Light)	2.76	9.64	20.76	19.54
S3: 3DUXNet	4.71	20.27	37.64	36.90
Aggregated Student performance derived from multiple Teachers (T1, T2, T3)				
H-Dense Former (Light)	1.25	3.31	72.37	71.39
Swin-UNETR (Light)	2.76	9.64	47.84	47.82
3DUXNet	4.71	20.27	43.15	42.44

Table 1. Comparison with different methods on HECKTOR21 and PI-CCAI22 test set.

publicly available source code on the two datasets. For the teacher variants of the models, we use a large number of features in the channel dimension to keep the architecture parameter-heavy and similar to the original works[4,38,39]. In contrast, for the student variants, we use a reduced number of features in channel dimensions to decrease the model's parameters $5\times$ to $10\times$ as seen in Table. 1. First, we train the three teacher models, H-Denseformer, SwinUNETR, and UNETR, successively on the HECKTOR21 and PI-CAI22 datasets. Next, we load the pre-trained weights of these three teacher models and freeze them. We then train our student network with the proposed knowledge distillation strategy in a 5-fold cross-validation setting. For MISSU and MDViT, we follow the training strategies outlined in their respective literature and utilize the official implementations provided by the authors. For MDViT, since we train on a single domain (one dataset), the domain-specific label is set to 1 during both training and inference. We also provide the 5-fold cross-validation result for our teacher, student and KD Models in the ablation study section in Table 2. In Figs. 3 and 4, we visualize segmentation results for the prediction of the heavy (teacher T1, T2, T3), lightweight (student S1, S2, S3),MISSU, MDVIT and proposed lightweight knowledge distillation-based model. It is apparent from the figure that our knowledge distillation approach generates a more accurate segmentation map than other lightweight student-only networks. Additionally, for MISSU, the output shape is irregular and does not match the rounded contours of the ground truth. In comparison, MDViT performs even worse, failing to generate any segmentation output for the given input.

Table 1 shows that our knowledge distillation-based approach has a superior dice score and mean-intersection-over-union (mIOU) compared to both teacher and student model variants. Generally, GFLOPS is used to check the throughput of the model's inference (time required), and the parameter count is to check the model's size (space required), which we provided in Table 1 for all the models. For HECKTOR21, 3D H-DenseFormer (lightweight), our knowledge-distillation approach achieves a dice score of 76.58%, and mIOU of 64.62%, which is a significant improvement over both teacher and student 3D H-DenseFormer[38], UNETR[4], and SwinUNETR[39] and student 3D UX-Net[12]. For PI-CAI22, the 2D variant of lightweight H-DenseFormer with our knowledge-distillation approach outperforms both teacher and student variants, and also against MDVIT by achieving a dice score of 72.37% and mIOU of 71.39%. Moreover, the number of parameters is almost $5\times$ to $10\times$, and GFLOPs are $10\times$ to $15\times$ smaller for our best-performing architecture (lightweight H-DenseFormer) on the two datasets. It should be mentioned that the higher the dice and mIOU score, the better, whereas the lower the parameters and GFLOPs, the better. To better illustrate the dice score vs. parameters and the mIOU vs. parameters, we showcase four plots in Fig. 1.

Ablation study

Cross-validation results

In Table 2, we provide the 5-fold cross-validation results with Dice score and mIOU for the three datasets in terms of mean and standard deviation. We choose the model with the best-performing validation score for testing.

HECKTOR (3D)				
Model	**Param(M)↓**	**GFLOPs↓**	**Dice↑**	**mIoU↑**
T1: H-DenseFormer	11.38	960.38	73.00 ± 2.72	61.60 ± 2.41
T2: Swin-UNETR	5.38	280.91	69.00 ± 1.67	56.99 ± 2.09
T3: UNETER	95.76	282.18	59.20 ± 3.24	47.40 ± 3.82
S1: H-Dense Former (Light)	1.53	62.73	64.20 ± 1.60	52.0 ± 1.67
S2: Swin-UNETR (Light)	1.35	75.68	63.0 ± 2.75	51.19 ± 2.40
S3: 3DUXNet	5.12	41.51	58.40 ± 2.33	46.00 ± 2.75
Aggregated Student performance derived from multiple Teachers (T1, T2, T3)				
H-Dense Former (Light)	1.53	62.73	73.60 ± 1.20	62.0 ± 1.09
Swin-UNETR (Light)	1.35	75.68	69.0 ± 1.67	56.99 ± 2.09
3DUXNet	5.12	41.51	69.19 ± 2.22	57.59 ± 1.95
PI-CAI (2D)				
T1: H-DenseFormer	4.98	32.33	45.00 ± 15.67	44.60 ± 16.26
T2: Swin-UNETR	11.04	37.94	33.40 ± 18.28	32.60 ± 19.22
T3: UNETER	87.51	59.42	13.99 ± 3.79	13.20 ± 3.18
S1: H-Dense Former (Light)	1.25	3.31	43.00 ± 9.38	42.60 ± 9.56
S2: Swin-UNETR (Light)	2.76	9.64	12.00 ± 6.35	12.00 ± 5.35
S3: 3DUXNet	4.71	20.27	21.80 ± 12.22	20.20 ± 13.05
Aggregated Student performance derived from multiple Teachers (T1, T2, T3)				
H-Dense Former (Light)	1.25	3.31	70.19 ± 2.20	70.19 ± 1.20
Swin-UNETR (Light)	2.76	9.64	21.60 ± 13.92	21.00 ± 14.35
3DUXNet	4.71	20.27	37.40 ± 13.09	36.80 ± 13.22

Table 2. Comparison with different methods on HECKTOR21[44] and PI-CCAI22[45] 5-fold cross-validation set. We provide the Dice and mIOU in terms of mean and standard deviation.

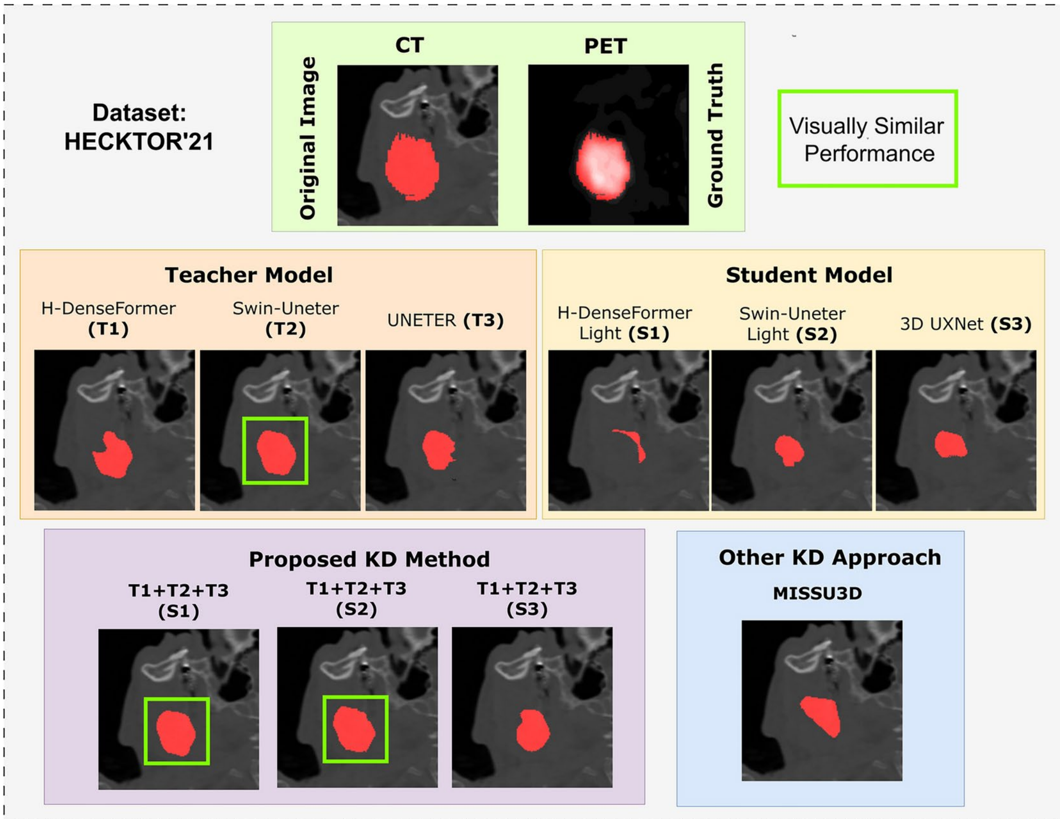

Fig. 3. Illustration of model's predictions on HECKTOR'21. The ground truth and prediction masks are overlapped on the original images (Red).

Effects of using different number of teachers

In the Table 3, we provide the effects of single, dual, and multi-teacher use on the best model's performance in terms of Dice score and mIoU. The table compares scenarios where the student model learns from a single teacher, pairs of teachers, and all three teachers combined. The results highlight that while each individual teacher contributes to the student's performance, using multiple teachers generally leads to improved segmentation accuracy. For instance, the student model trained with all three teachers (T1+T2+T3) achieves the highest Dice score and mIoU, suggesting that the combined knowledge from multiple teachers provides a more comprehensive understanding, leading to more accurate segmentation. This finding emphasizes the effectiveness of multi-teacher knowledge distillation in enhancing the performance of lightweight student models, making them more robust and capable in comparison to those trained with a single teacher.

Effects of knowledge distillation loss

In Table 4, we have provided the effects of loss functions, namely, $\mathcal{L}_{\mathcal{AFS}}$ and $\mathcal{L}_{\mathcal{PWD}}$ on the Dice score and mIOU of the best performing model (H-DenseFormer light trained with knowledge distillation) on the HECKTOR21 dataset. The table compares the outcomes when applying individual loss functions-coarse and fine attention feature similarity loss $\mathcal{L}_{\mathcal{AFS}}$ and pixel-wise distillation loss $\mathcal{L}_{\mathcal{PWD}}$ -as well as the combined effect of both losses. The data clearly shows that each loss function independently contributes to the improvement of the student model's performance. However, the combination of $\mathcal{L}_{\mathcal{AFS}}$ and $\mathcal{L}_{\mathcal{PWD}}$ results in the most substantial gains in both Dice score and mIoU. This combined approach enables the student model to effectively capture both the spatial relationships and detailed pixel-level information from the teacher models, leading to a more accurate and robust segmentation. The results highlight the effectiveness of using a multifaceted distillation strategy, where integrating different loss components allows for a more comprehensive transfer of knowledge from the teacher models.

Statistical significance

To further validate the effectiveness of our proposed model, we conducted paired t-tests. These tests aimed to determine the statistical significance of the performance differences between our, the teacher and the student models for H-DenseFormer on the HECKTOR21 and PI-CAI22 datasets. We chose H-DenseFormer as a base comparison as it outperformed other lightweight models. Our approach to statistical significance testing is crucial for ensuring that the observed improvements in performance are not merely due to random variations (null hypothesis) in the data but represent a genuine advancement in model accuracy.

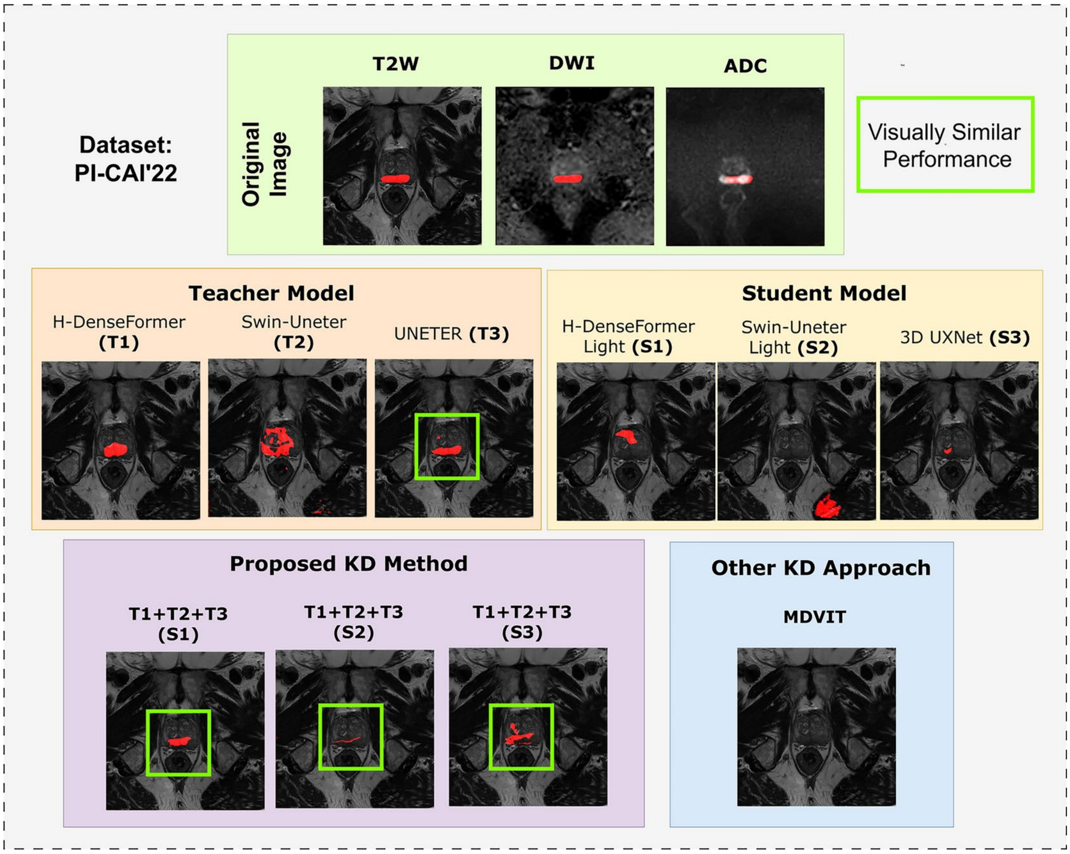

Fig. 4. Illustration of model's predictions on PI-CAI22. The ground truth and prediction masks are overlapped on the original images (Red).

H-DenseFormer (T1)	SwinUNETER (T2)	UNETER (T3)	Dice	mIoU
✓			75.88	64.07
	✓		76.15	65.12
		✓	73.26	62.24
✓	✓		74.72	62.82
	✓	✓	76.75	64.86
✓		✓	75.37	63.42
✓	✓	✓	76.58	64.62

Table 3. Effects of using single, dual, and multi teacher on the best-performing model's (H-Denseformer) performance on HECKTOR21 dataset[44].

$\mathscr{L}_{AFS}$	$\mathscr{L}_{PWD}$	Dice	mIoU
✓		73.27	60.66
	✓	74.12	61.80
✓	✓	76.58	64.62

Table 4. Effects of various knowledge distillation loss on the best-performing model's (H-Denseformer) performance on HECKTOR21 dataset.

For the HECKTOR21 dataset, the comparison between our model and the H-DenseFormer teacher resulted in a *p*-value of 0.025, which is *leq*0.05, indicating the model's prediction is statistically significant and good evidence against the null hypothesis. Furthermore, the comparison between our model and the H-DenseFormer student yielded an even more significant *p*-value of 0.0004, which is *leq*0.001, demonstrating that our model has

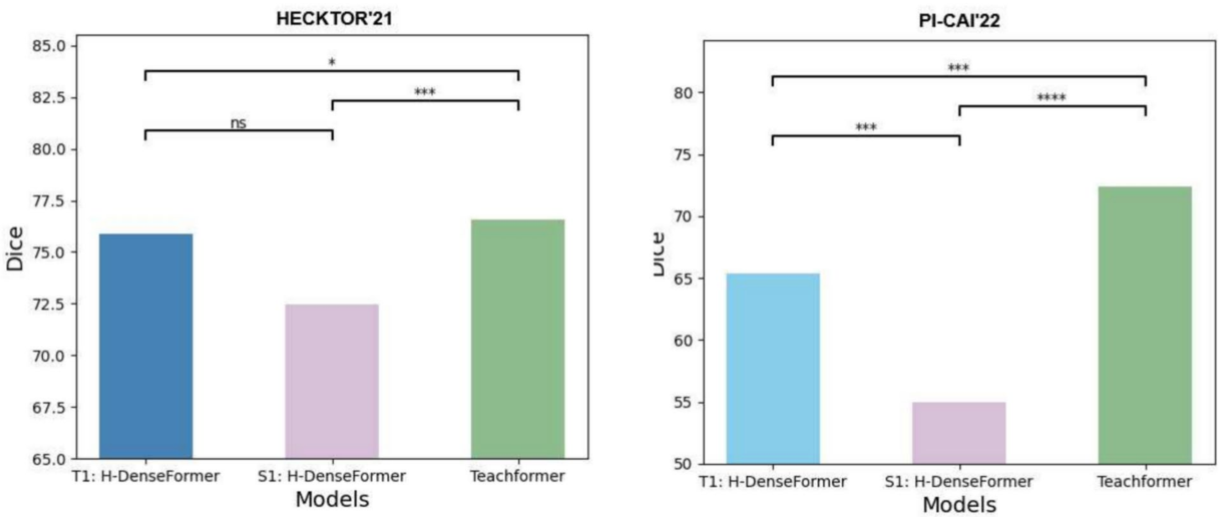

Fig. 5. This figure presents the p-values from statistical significance tests comparing the proposed model against both teacher (T) and student (S) models on the HECKTOR21 and PI-CAI22 datasets. The X-axis and Y-axis signify models and dice-score successively. Significance levels are indicated by asterisks: $(*)$ $p \leq 0.05$, $(***)$ $p \leq 0.001$, $(****)$ $p \leq 0.0001$., and n.s. $p > 0.05$..

strong evidence against the null hypothesis. The *p*-value for the teacher and student model was > 0.05, meaning the teacher model had no significant (n.s) improvement over the student model.

Similarly, for the PI-CAI22 dataset, our model showed very strong statistically significant improvements over both the teacher and student models. The comparison with the H-DenseFormer teacher resulted in a p-value of 0.0003, which is *leq*0.001, while the comparison with the H-DenseFormer student produced a p-value of 0.00007, which is *leq*0.0001. These extremely low p-values indicate strong evidence against the teacher model and strong evidence against the student model for the null hypothesis. When we compare the teacher and student models for the same samples for PI-CAI22, we see that the p-value is *leq*0.001, meaning that the teacher has a significant performance compared to the student model.

The consistent statistical significance across both datasets underscores the robustness and reliability of our model. Figure 5 demonstrates that consistently achieving p-values well below the conventional threshold of 0.05 confirms that our model's superior performance in terms of Dice score is statistically validated and practically meaningful. The proposed Teach-Former model offers significant improvements over both the teacher and student versions of H-DenseFormer, reinforcing its potential for real-world clinical applications in medical image segmentation.

Conclusion

We introduce "Teach-Former," a transformative strategy for medical image segmentation that leverages a Transformer-based knowledge distillation framework to consolidate the strengths of multiple teacher models into a single, efficient, and lightweight student model. This method demonstrates significant advantages in processing multimodal imaging data (CT, PET, MRI) with reduced computational demands, offering enhanced performance and efficiency, as evidenced by evaluations on the HECKTOR21 and PI-CAI22 datasets. The Teach-Former approach holds promise in advancing clinical care and precision medicine by utilizing multimodal data to improve segmentation accuracy, which may lead to more precise diagnoses and better-informed treatment planning. Its deep learning algorithms are adept at distinguishing and delineating anatomical structures and pathological features across different imaging modalities, providing a comprehensive view of the patient's health status and identifying nuanced patterns that may be overlooked with single-modality approaches.

However, the success of Teach-Former is contingent on the quality and completeness of the multimodal data it processes. Inconsistencies, incomplete datasets, or noisy clinical data can potentially undermine the model's predictive accuracy and segmentation precision. Despite these challenges, Teach-Former has the potential to improve diagnostic accuracy, streamline clinical decision-making, and enhance workflow efficiency. Future work will further validate the strategy and expand its application in real-world clinical settings to continuously refine the model based on new data and medical advancements to better diagnose and manage various health conditions.

Data availibility

The HECKTOR'21 dataset was released as part of the second edition of the HEad and neCK TumOR (HECKTOR) challenge, organized as a satellite event of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2021. The PI-CAI dataset was published as a Grand Challenge to identify tumors from prostate MRI exams. Data is provided within the manuscript. The dataset is public. Affiliate references are cited in the manuscript.

Received: 8 October 2024; Accepted: 20 February 2025
Published online: 07 May 2025

References

1. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, 234–241 (Springer, 2015).
2. Jin, Q., Meng, Z., Sun, C., Cui, H. & Su, R. RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans. *Front. Bioeng. Biotechnol.* **8**, 605132 (2020).
3. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
4. Hatamizadeh, A. *et al.* Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574–584 (2022).
5. Cao, H. *et al.* Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, 205–218 (Springer, 2022).
6. Zhao, L. et al. MSKD: Structured knowledge distillation for efficient medical image segmentation. *Comput. Biol. Med.* **164**, 107284 (2023).
7. Park, S. et al. Self-evolving vision transformer for chest x-ray diagnosis through knowledge distillation. *Nat. Commun.* **13**, 3848 (2022).
8. Liao, Z. *et al.* Knowledge distillation of attention and residual u-net: Transfer from deep to shallow models for medical image classification. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 162–173 (Springer, 2023).
9. Andrade-Miranda, G. et al. Multi-modal medical transformers: A meta-analysis for medical image segmentation in oncology. *Comput. Med. Imaging Graph.* **110**, 102308 (2023).
10. Marias, K. The constantly evolving role of medical image processing in oncology: From traditional medical image processing to imaging biomarkers and radiomics. *J. Imaging* **7**, 124 (2021).
11. Zhang, Y., Shen, Z. & Jiao, R. Segment anything model for medical image segmentation: Current applications and future directions. *Comput. Biol. Med.* **171**, 108238 (2024).
12. Lee, H. H., Bao, S., Huo, Y. & Landman, B. A. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. arXiv preprint arXiv:2209.15076 (2022).
13. Liu, C. & Kiryu, H. 3d medical axial transformer: A lightweight transformer model for 3d brain tumor segmentation. In *Medical Imaging with Deep Learning*, 799–813 (PMLR, 2024).
14. Du, X. et al. Al-net: Asymmetric lightweight network for medical image segmentation. *Front. Signal Process.* **2**, 842925 (2022).
15. Polino, A., Pascanu, R. & Alistarh, D. Model compression via distillation and quantization. arXiv preprint arXiv:1802.05668 (2018).
16. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems***32** (2019).
17. Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015).
18. Qin, D. et al. Efficient medical image segmentation based on knowledge distillation. *IEEE Trans. Med. Imaging* **40**, 3820–3831 (2021).
19. Pham, C., Hoang, T. & Do, T.-T. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6435–6443 (2023).
20. Dong, A. et al. Momentum contrast transformer for COVID-19 diagnosis with knowledge distillation. *Pattern Recogn.* **143**, 109732 (2023).
21. Kanwal, N., Eftestøl, T., Khoraminia, F., Zuiverloon, T. C. & Engan, K. Vision transformers for small histological datasets learned through knowledge distillation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 167–179 (Springer, 2023).
22. Huang, X., Deng, Z., Li, D. & Yuan, X. Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021).
23. Zhong, Y., Yang, Y., Fang, Y., Wang, J. & Hu, W. A preliminary experience of implementing deep-learning based auto-segmentation in head and neck cancer: A study on real-world clinical cases. *Front. Oncol.* **11**, 638197 (2021).
24. Kosmin, M. et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother. Oncol.* **135**, 130–140 (2019).
25. Kitaguchi, D. et al. Computer-assisted real-time automatic prostate segmentation during TaTME: A single-center feasibility study. *Surg. Endosc.* **35**, 2493–2499 (2021).
26. Liang, P., Chen, J., Chang, Q. & Yao, L. Rskd: Enhanced medical image segmentation via multi-layer, rank-sensitive knowledge distillation in vision transformer models. *Knowl.-Based Syst.* **293**, 111664 (2024).
27. Valindria, V. V. *et al.* Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 547–556 (IEEE, 2018).
28. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
29. Martí-Bonmatí, L., Sopena, R., Bartumeus, P. & Sopena, P. Multimodality imaging techniques. *Contrast Media Mol. Imaging* **5**, 180–189 (2010).
30. Guo, Z., Li, X., Huang, H., Guo, N. & Li, Q. Deep learning-based image segmentation on multimodal medical imaging. *IEEE Trans. Radiat. Plasma Med. Sci.* **3**, 162–169 (2019).
31. Li, X., Chen, S., Hu, X. & Yang, J. Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2682–2690 (2019).
32. Song, Y. et al. Medical image classification: Knowledge transfer via residual u-net and vision transformer-based teacher-student model with knowledge distillation. *J. Vis. Commun. Image Represent.* **102**, 104212 (2024).
33. Du, S., Bayasi, N., Hamarneh, G. & Garbi, R. Mdvit: Multi-domain vision transformer for small medical image segmentation datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 448–458 (Springer, 2023).
34. Wang, N. MISSU: 3D medical image segmentation via self-distilling TransUNet. *IEEE Trans. Med. Imaging* **42**(9), 2740–2750 (2023).
35. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571 (IEEE, 2016).
36. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19*, 424–432 (Springer, 2016).
37. Isensee, F. *et al.* nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018).

38. Shi, J. *et al.* H-denseformer: An efficient hybrid densely connected transformer for multimodal tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 692–702 (Springer, 2023).
39. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, 272–284 (Springer, 2021).
40. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).
41. Zagoruyko, S. & Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016).
42. He, T. *et al.* Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 578–587 (2019).
43. Liu, Y. *et al.* Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2604–2613 (2019).
44. Andrearczyk, V. *et al.* Overview of the HECKTOR challenge at MICCAI 2021: Automatic head and neck tumor segmentation and outcome prediction in PET/CT images. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, 1–37 (Springer, 2021).
45. Saha, A. *et al.* The PI-CAI challenge: Public training and development dataset (2022).
46. Yaqub, M. et al. State-of-the-art CNN optimizer for brain tumor segmentation in magnetic resonance images. *Brain Sci.* **10**, 427 (2020).
47. Feng, C.-M. Enhancing label-efficient medical image segmentation with text-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 253–262 (Springer, 2024).
48. Yao, W. et al. From cnn to transformer: A review of medical image segmentation models. *J. Imaging Inf. Med.* **37**(4), 1529–1547 (2024).

Acknowledgements
This work was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under grant number P30 *GM*145646 and by the National Science Foundation under grant number OAC 2201599 and grant number OIA 2148788.

Author contributions
Literature search (K.F.H). Conception and design (K.F.H) Statistical expertise (K.F.H and S.A.K) Analysis and interpretation (K.F.H and S.A.K.) Writing the article (K.F.H., S.A.K., and J.O.) Critical revision (A.T. and J.O.); Final approval of the article (A.T., S.A.K., K.F.H, and J.O.);

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to K.F.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.