Task-Performance Routing for Multi-Teacher Distillation in CT Medical Image Segmentation

Jiaye Yang[1], Yao Liu[1(✉)], Die Dai[1], Hanwen Zhang[1], Peiyuan Jiang[1], Qiao Liu[1], Yutong Xie[2], and Peng Wang[1]

[1] University of Electronic Science and Technology of China, Chengdu, 611731, China
[2] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
liuyao@uestc.edu.cn

Abstract. Large-scale pre-trained models have achieved remarkable performance in CT organ segmentation, yet individual models exhibit complementary strengths across different anatomical regions. While multi-teacher knowledge distillation can leverage these complementary advantages, existing methods assign teacher weights at the instance or global level, overlooking the inherent spatial heterogeneity of CT volumes where tissue contrast, boundary ambiguity, and organ morphology vary significantly across regions. We propose Task-Performance Routing (TPR), a novel framework that dynamically routes to the best-performing teacher at each spatial region based on actual prediction errors rather than learned parameters. TPR employs hierarchical spatial adaptation—fine-grained routing in shallow layers for boundary details and coarse-grained routing in deep layers for semantic coherence—and applies performance-based selection in hard regions while encouraging complementary learning in easy regions. Extensive experiments on BTCV, MSD-Liver, and MSD-Pancreas datasets demonstrate that TPR consistently outperforms individual teachers, adaptive ensemble methods, and existing multi-teacher distillation baselines, achieving superior segmentation accuracy without inference overhead. Code is available at https://github.com/EdvinCecilia/TPR_MultiTeacher.

Keywords: Multi-Teacher Distillation · CT Organ Segmentation · Foundation Models

1 Introduction

Large-scale pre-trained models have significantly advanced CT organ segmentation, yet different pre-training paradigms (e.g., SuPreM [11], VoCo [22], and CLIP-Driven [13]) yield complementary representations, excelling in different organs and spatial regions due to their distinct training objectives and data distributions. We analyze SuPreM and VoCo pre-training representations on multi-organ segmentation data BTCV [10]. Despite similar overall Dice scores, they exhibit distinct spatial activation patterns and complementary organ-level performance (Fig. 1), suggesting that spatially adaptive routing can exploit these differences more effectively than global weighting schemes.

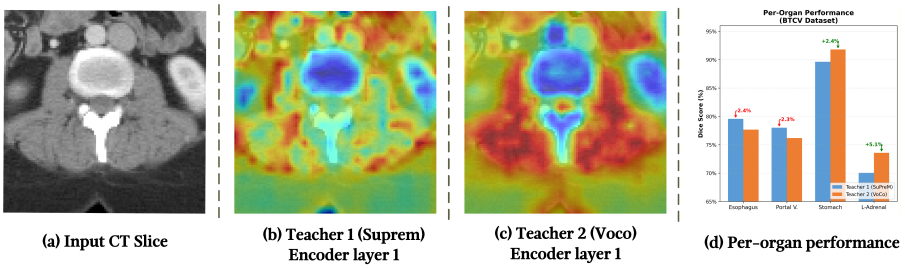

Fig. 1. Motivation on BTCV: Complementary performance of two teachers. **(a)** Input CT slice. **(b–c)** Encoder feature activation maps from SuPreM and VoCo reveal distinct spatial focus: SuPreM activates strongly around small tubular structures, whereas VoCo responds more broadly along large-organ boundaries. **(d)** Per-organ Dice scores confirm organ-wise complementarity: SuPreM (blue) excels at the esophagus and portal vein, whereas VoCo (orange) is stronger for the stomach and left adrenal gland.

Multi-teacher knowledge distillation (MTKD) [6,4] offers a natural way to merge these strengths into a single student without inference overhead. Existing MTKD methods have evolved from uniform weighting [7] to instance-level adaptive schemes based on latent representations [14,24] or gradient-space optimization [3], as well as multi-scale self-distillation strategies [9]. In medical imaging, MTKD has been primarily explored for cross-modal distillation [1,8], while intra-modality settings remain underexplored [21]. Furthermore, prior work typically trains teachers from scratch rather than distilling from independently pretrained foundation models—a setting in which the goal shifts from compression to knowledge fusion from independently pre-trained foundation models [16]. Crucially, none of these approaches exploit region-wise spatial heterogeneity within 3D CT volumes, where tissue contrast, boundary complexity, and organ morphology vary across spatial locations within the same sample.

To address this gap, we propose Task-Performance Routing (TPR), which dynamically assigns region-wise teacher weights based on actual segmentation errorsselecting the best-performing teacher in hard regions while encouraging complementary fusion in easy regions. This performance-driven routing operates at hierarchical spatial granularity, with fine-grained routing in shallow layers for boundary details and coarse-grained routing in deep layers for semantic coherence, requiring no learned routing parameters while naturally adapting to the spatial heterogeneity of CT volumes. Experiments on BTCV, MSD-Liver, and MSD-Pancreas with different teacher pairs show that TPR consistently outperforms individual teachers, adaptive ensemble methods, and existing MTKD baselines. Our main contributions are:

– **Task-performance routing:** region-wise teacher selection driven by prediction error with hierarchical spatial granularity, requiring no extra routing parameters.

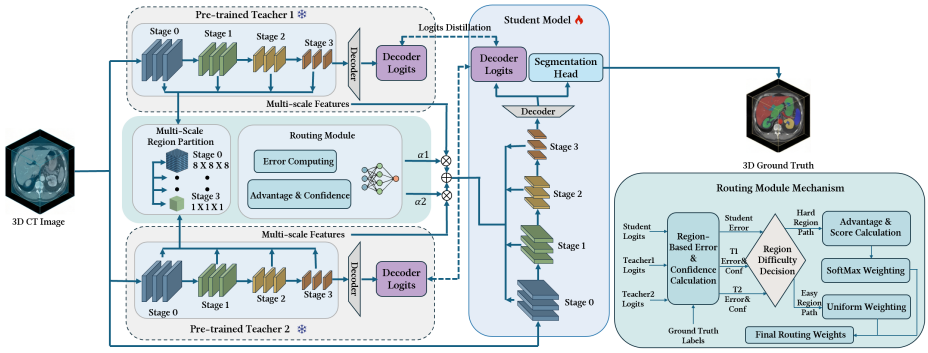

Fig. 2. Overall framework of Task-Performance Routing (TPR).

– **Complementarity-aware distillation:** effective knowledge fusion from independently pre-trained foundation models that surpasses both individual teachers and ensemble baselines without inference overhead.

2 Method

Figure 2 provides an overview of TPR, illustrating the hierarchical routing mechanism and the feature alignment process across encoder stages.

2.1 Overview and Problem Formulation

Given an input CT volume $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$ and its ground truth segmentation mask $\mathbf{y} \in \{0, 1, ..., C\}^{H\times W\times D}$ with C organ classes, we aim to train a student model $\mathcal{S}_\theta$ that learns from multiple frozen pre-trained teacher models $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_M\}$, each independently pre-trained under heterogeneous paradigms. Our goal is to enable the student to surpass the performance of individual teachers by effectively combining their complementary strengths across different spatial regions.

Both teachers and student share an identical encoder-decoder architecture: SwinUNETR [5] for the VoCo+SuPreM pair and U-Net [18] for the SuPreM+CLIP pair, extracting multi-scale features at L stages. For the student, we denote the encoder features at stage l as $\mathbf{f}_s^l \in \mathbb{R}^{C_l\times H_l\times W_l\times D_l}$ and the final output logits as $\mathbf{L}_s \in \mathbb{R}^{C\times H\times W\times D}$. Similarly, teacher m produces features $\mathbf{f}_{t_m}^l$ and logits $\mathbf{L}_{t_m}$. All teacher models are completely frozen during training, requiring no gradient computation. We propose TPR, which computes adaptive routing weights $\mathbf{w}_m^{l,r}$ for teacher m at stage l and region r based on actual segmentation performance, in contrast to traditional MTKD methods that rely on spatially uniform or globally adaptive weights.

2.2 Task-Performance Router

The TPR dynamically determines teacher weights based on actual prediction errors, comprising three components: hierarchical region division, routing weight generation, and hard/easy region adaptation.

Hierarchical Region Division Medical image segmentation involves features at different abstraction levels—shallow layers capture fine-grained boundary details while deep layers encode global semantic context. To match routing granularity with feature hierarchy, we apply fixed 3D grid division at each encoder stage l.

Given a feature map $\mathbf{f}^l \in \mathbb{R}^{C_l \times H_l \times W_l \times D_l}$ at stage l, we divide it into N_l non-overlapping 3D grid regions. The region size R_l decreases from shallow to deep layers:

$$R_l = \begin{cases} 8 \times 8 \times 8, & l = 0 \text{ (shallow)} \\ 4 \times 4 \times 4, & l = 1 \\ 2 \times 2 \times 2, & l = 2 \\ 1 \times 1 \times 1, & l = 3 \text{ (deep)} \end{cases} \tag{1}$$

This hierarchical design reflects the intuition that shallow layers require fine-grained spatial routing to handle local variations in tissue contrast and boundary clarity, while deep layers benefit from coarse or global routing to maintain semantic coherence across organs.

Routing Weight Generation We first compute a voxel-wise cross-entropy error map from the final logits of the student and each frozen teacher, then pool this map using the stage-specific 3D grid to obtain region-level errors e_s^r and $e_{t_m}^r$ at each encoder stage. Regions with larger student error are treated as hard regions ($e_s^r > \tau_{\text{hard}}$, where τ_{hard} is a threshold hyperparameter), motivated by the observation that hard examples contribute disproportionately to learning [12]. This hard/easy region distinction also draws on curriculum learning principles [2], where harder examples receive more focused supervision. In hard regions, we select the best teacher via softmax over each teacher's relative advantage; in easy regions, we apply uniform weights to encourage the student to absorb complementary knowledge from all teachers rather than overfitting to a single one:

$$w_m^{l,r} = \begin{cases} \dfrac{\exp\left(\frac{e_s^r - e_{t_m}^r}{\tau}\right)}{\sum_{m'=1}^{M} \exp\left(\frac{e_s^r - e_{t_{m'}}^r}{\tau}\right)}, & e_s^r > \tau_{\text{hard}} \\ \dfrac{1}{M}, & \text{otherwise} \end{cases} \tag{2}$$

2.3 Multi-Component Loss Function

The total training objective combines four terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{logits}}\mathcal{L}_{\text{logits}} + \lambda_{\text{balance}}\mathcal{L}_{\text{balance}} \tag{3}$$

Segmentation loss provides direct supervision from ground truth using a combination of Dice and cross-entropy losses: $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}}(\mathbf{L}_s, \mathbf{y}) + \mathcal{L}_{\text{CE}}(\mathbf{L}_s, \mathbf{y})$.
Feature alignment loss aligns student encoder features with teacher features using region-adaptive routing weights across all $L = 4$ encoder stages, from fine-grained shallow features to coarse semantic deep features. Here $\bar{\mathbf{f}}^{l,r}$ denotes the spatially averaged feature vector within region r at stage l:

$$\mathcal{L}_{\text{align}} = \sum_{l=0}^{L-1} \frac{1}{N_l} \sum_{r=1}^{N_l} \sum_{m=1}^{M} w_m^{l,r} \left\| \bar{\mathbf{f}}_s^{l,r} - \bar{\mathbf{f}}_{t_m}^{l,r} \right\|_2^2 \tag{4}$$

Logits distillation loss transfers knowledge from teachers' final predictions via KL divergence [17] on region-averaged logits, where $\bar{\mathbf{L}}^r$ denotes the spatially averaged logit vector within region r:

$$\mathcal{L}_{\text{logits}} = \frac{1}{N_{\text{out}}} \sum_{r=1}^{N_{\text{out}}} \sum_{m=1}^{M} w_m^{L,r} \text{KL}\left(p_s(\bar{\mathbf{L}}_s^r) \| p_{t_m}(\bar{\mathbf{L}}_{t_m}^r)\right) \tag{5}$$

Load balance regularization prevents routing collapse, inspired by load balancing in mixture-of-experts models [19]. We maximize the entropy of average routing weights $\bar{\mathbf{w}}$:

$$\mathcal{L}_{\text{balance}} = \exp\left(-\frac{\mathcal{H}(\bar{\mathbf{w}})}{\tau_b}\right), \quad \bar{w}_m = \frac{1}{N_{\text{total}}} \sum_{l,r} w_m^{l,r} \tag{6}$$

where $\mathcal{H}(\bar{\mathbf{w}}) = -\sum_{m=1}^{M} \bar{w}_m \log \bar{w}_m$ and $\tau_b = 0.4$. We set $\lambda_{\text{seg}} = 1.0$, $\lambda_{\text{align}} = 0.2$, $\lambda_{\text{logits}} = 0.2$, $\lambda_{\text{balance}} = 0.1$.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate TPR on BTCV [10] and two MSD tasks [20] (Liver and Pancreas). For BTCV, we follow the commonly-used split (24 training / 6 testing) [5]. For MSD, we use the official training set provided by the challenge and create a held-out test split (80% training / 20% testing, random seed 42) within it for all comparisons. All volumes are resampled to $1.5 \times 1.5 \times 2.0$ mm and intensities are clipped to $[-175, 250]$ HU. We report mean Dice (%, higher is better) and mean HD95 (mm, lower is better) over all foreground classes.

Table 1. Comparison of TPR student with individual teachers and multi-teacher distillation baselines on three datasets. We report Dice (%, higher is better) and HD95 (mm, lower is better). Best results are **bold**, second best underlined.

		SwinUNETR-based *(VoCo + SuPreM)*			**U-Net-based** *(SuPreM + CLIP)*		
Method	**Metric**	BTCV	Liver	Panc.	BTCV	Liver	Panc.
T1	Dice (%)	85.09	81.42	62.28	83.40	82.20	65.60
	HD95	2.57	14.65	15.33	2.68	13.95	16.12
T2	Dice (%)	84.50	81.79	63.72	84.20	83.00	67.37
	HD95	2.71	13.82	17.99	2.55	13.45	15.68
Ensemble (adaptive)	Dice (%)	85.15	82.28	64.38	84.68	83.78	67.45
	HD95	2.50	13.48	15.62	2.62	13.48	15.92
AE-KD [3] (NIPS'20)	Dice (%)	85.10	81.85	63.58	84.38	83.22	66.72
	HD95	2.65	13.88	16.52	2.68	13.72	16.05
MMKD [25] (ICME'23)	Dice (%)	85.25	81.98	63.89	84.52	83.58	67.12
	HD95	2.58	13.65	16.18	2.64	13.55	15.98
MTKD-RL [23] (AAAI'25)	Dice (%)	85.23	82.15	64.52	84.75	84.08	67.58
	HD95	2.54	13.38	15.78	2.58	13.32	15.72
TPR (Ours)	Dice (%)	**85.45**	**82.80**	**65.40**	**85.20**	**84.70**	**68.25**
	HD95	**2.40**	**13.08**	**15.24**	**2.50**	**13.15**	**15.45**

3.2 Teacher Models and Experimental Settings

We use two teacher pairs: (VoCo [22], SuPreM [11]) with SwinUNETR and (SuPreM [11], CLIP-Driven [13]) with U-Net; the student shares the same architecture as its teachers to enable feature alignment. All models use a patch size of $96 \times 96 \times 96$. Since teacher models only release encoder weights, we train teacher decoders from scratch for 1200 epochs with frozen encoders. Students are trained for 1200 epochs using AdamW [15] (lr=3×10^{-4}, weight decay=10^{-5}, 20 warmup epochs), with $\tau_{\text{hard}} = 0.5$ and $\tau = 0.7$. Data augmentation includes random flipping and rotation (p=0.2). All experiments use 4 NVIDIA RTX 4090 GPUs. The adaptive ensemble baseline applies softmax-normalized weights based on each teacher's validation Dice score.

3.3 Comparison Study

As shown in Table 1, TPR consistently outperforms both individual teachers, adaptive ensemble, and all MTKD baselines across three datasets and both teacher pair settings. Notably, TPR surpasses even the adaptive ensemble method, demonstrating that our distillation approach achieves superior knowledge fusion rather than simply aggregating teacher outputs.

Table 2. Component ablation on MSD-Pancreas (SwinUNETR-based setting).

Variant	Mean Dice (%)
Teacher 1 (T1)	62.3
Teacher 2 (T2)	63.7
Baseline (uniform multi-teacher KD)	63.9
w/o global performance weighting	64.3
w/o spatial routing (hard/easy split)	64.8
w/o hierarchical granularity	65.1
TPR (full)	**65.4**

Table 3. Routing granularity strategies on MSD-Pancreas.

Routing Strategy	Mean Dice (%)
All fine-grained ($8 \times 8 \times 8$ at all stages)	64.6
All coarse-grained ($1 \times 1 \times 1$ at all stages)	64.2
Uniform granularity ($4 \times 4 \times 4$ at all stages)	64.9
Hierarchical (ours)	**65.4**

Table 4. Loss component ablation on MSD-Pancreas (SwinUNETR-based, 1200 epochs).

$\mathcal{L}_{align}$	$\mathcal{L}_{logits}$	$\mathcal{L}_{balance}$	Mean Dice (%)
✓	–	✓	64.52
–	✓	✓	64.78
✓	✓	–	65.11
✓	✓	✓	**65.40**

3.4 Ablation Studies

We conduct ablations in the SwinUNETR-based setting on MSD-Pancreas to validate the key design choices of TPR. All variants are trained for 1200 epochs under identical settings.

As shown in Tables 2–4, each design choice contributes consistently and statistically significantly ($p < 0.05$, paired t-test).

Table 2 shows the incremental contribution of each routing component. Global performance weighting lifts performance by 0.4% over the uniform baseline. Spatial routing (hard/easy split) adds a further 0.5%, demonstrating that region-wise teacher selection in difficult areas is more effective than globally adaptive weighting. Hierarchical granularity contributes an additional 0.3%, confirming that matching routing resolution to feature abstraction level further improves knowledge transfer. $\tau_{\text{hard}} = 0.5$ corresponds to regions exceeding the dataset mean cross-entropy error, and yields stable performance across $[0.4, 0.6]$.

Table 3 isolates routing granularity strategies. Fine-grained routing at deep layers introduces spatial noise that disrupts semantic coherence, while coarse routing at shallow layers fails to capture local boundary variations. Uniform granularity ($4 \times 4 \times 4$) falls short of hierarchical design by 0.5%.

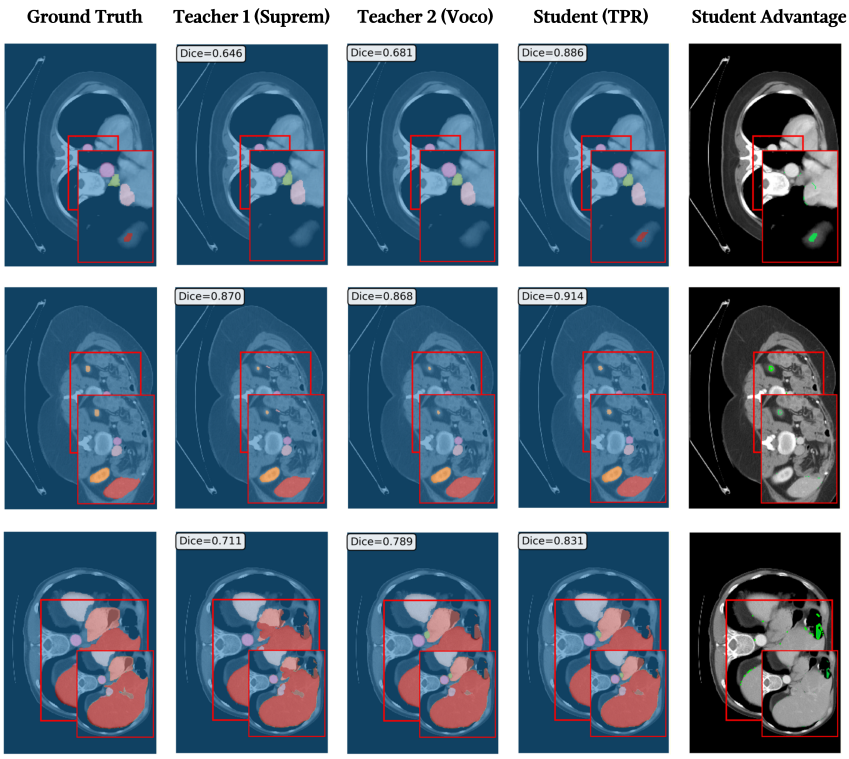

Fig. 3. Qualitative comparison on BTCV.

Table 4 shows all three loss components are complementary. Removing $\mathcal{L}_{align}$ causes the largest drop (0.88%), confirming that feature-level alignment provides richer supervision signals than output-level distillation alone in our frozen-teacher setting. Removing $\mathcal{L}_{logits}$ leads to a 0.62% drop, while removing $\mathcal{L}_{balance}$ causes routing collapse toward a single dominant teacher, hurting generalization by 0.29%.

3.5 Qualitative Results

Figure 3 presents representative axial slices from the BTCV dataset, comparing segmentation outputs of both teachers and our student against the ground truth. The rightmost column visualizes the *Student Advantage* map, highlighting regions where the student outperforms both teachers simultaneously.

Teacher 1 (SuPreM) and Teacher 2 (VoCo) exhibit complementary failure patterns: SuPreM tends to over-segment neighboring organs, while VoCo frequently misses fine-grained boundaries in cluttered regions. Despite being distilled from these imperfect teachers, TPR produces boundaries that more closely align with the ground truth, as shown in the zoomed-in red box regions. The

Student Advantage maps confirm that these improvements emerge from the distillation process itself rather than being inherited from either teacher.

4 Conclusion

We propose Task-Performance Routing (TPR) for multi-teacher distillation in CT segmentation, assigning region-wise teacher weights based on task error at hierarchical spatial granularity. TPR effectively exploits complementary strengths of different foundation models and yields consistent improvements over individual teachers. One limitation is that TPR requires ground truth labels for error-based routing, making it less suitable for semi-supervised or label-scarce settings; pseudo-label-based routing remains future work.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (U22B2061), the National Key R&D Program of China (2022YFB4300603), the Natural Science Foundation of Sichuan, China (2026NSFSC0429), and the National Natural Science Foundation of China under Grant No. 62572100.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmad, S., Ullah, Z., Gwak, J.: Multi-teacher cross-modal distillation with cooperative deep supervision fusion learning for unimodal segmentation. Knowledge-Based Systems **297**, 111854 (2024)
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
3. Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., Zhang, C.: Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. advances in neural information processing systems **33**, 12345–12355 (2020)
4. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International journal of computer vision **129**(6), 1789–1819 (2021)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
6. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat **1050**, 9 (2015)
7. Hossain, K.F., Kamran, S.A., Ong, J., Tavakkoli, A.: Enhancing efficient deep learning models with multimodal, multi-teacher insights for medical image segmentation. Scientific Reports **15**(1), 15948 (2025)
8. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 772–781. Springer (2020)
9. Jia, Z., Sun, S., Liu, G., Liu, B.: Mssd: multi-scale self-distillation for object detection. Visual Intelligence **2**(1), 8 (2024)

10. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI multi-atlas labeling beyond cranial vaultworkshop challenge. vol. 5, p. 12. Munich, Germany (2015)
11. Li, W., Yuille, A., Zhou, Z.: How well do supervised models transfer to 3d image segmentation. In: The Twelfth International Conference on Learning Representations. vol. 1 (2024)
12. 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. pp. 2980–2988 (2017)
13. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21152–21164 (2023)
14. Liu, Y., Zhang, W., Wang, J.: Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing **415**, 106–113 (2020)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
17. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
19. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
20. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms (2019)
21. Wang, Y., Cao, P., Hou, Q., Lan, L., Yang, J., Liu, X., Zaiane, O.R.: Progressively correcting soft labels via teacher team for knowledge distillation in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 521–530. Springer (2024)
22. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22873–22882 (2024)
23. Yang, C., Yu, X., Yang, H., An, Z., Yu, C., Huang, L., Xu, Y.: Multi-teacher knowledge distillation with reinforcement learning for visual recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9148–9156 (2025)
24. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: International Conference on Learning Representations (2017)
25. Zhang, H., Chen, D., Wang, C.: Adaptive multi-teacher knowledge distillation with meta-learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 1943–1948. IEEE (2023)