Displacement Preserving Relational Distillation for Robust Medical Segmentation

Zhicheng Ding[1,3*], Xinyu Chu[1,3*], Jung Im Choi[2], Qing Tian[3†], Tianyu Shi[4], Xiaoqian Jiang[6], Lijing Zhu[5], and Qizhen Lan[6*†]

[1] Bowling Green State University, Bowling Green, OH, USA
[2] The University of Findlay, Findlay, OH, USA
[3] University of Alabama at Birmingham, Birmingham, AL, USA
qtian@uab.edu
[4] University of Toronto, Toronto, Ontario, Canada
[5] University of Houston - Clear Lake, Houston, TX, USA
[6] UTHealth Houston, Houston, TX, USA
qizhen.lan@uth.tmc.edu

Abstract. Accurate 3D medical segmentation requires models that are both robust to anatomical variability and efficient for deployment. While knowledge distillation (KD) supports model compression, conventional voxel-wise or scalar-relation methods can be sensitive to architectural mismatch and background-dominated 3D volumes. We propose Displacement Preserving Relational Distillation (DPRD), which aligns normalized pairwise displacement vectors among ROI-pooled case-level embeddings to provide scale-invariant local relational supervision. Integrated into nnU-Net, DPRD outperforms representative KD baselines on ISLES 2022 and AMOS 2022. On AMOS, DPRD achieves 85.46% Dice and 11.72 mm HD95, slightly exceeding the MedNeXt teacher in Dice while the MobileUNetV3 student uses only ∼5% of its parameters and ∼3% of its FLOPs. Code: https://github.com/ClinicaAlpha/DPRD-3D-MedSeg

Keywords: 3D Medical Image Segmentation · Knowledge Distillation

1 Introduction

Clinicians must precisely localize 3D operative targets and at-risk structures during image-guided procedures, often under strict workflow constraints [11,17,19]. This necessitates reliable 3D segmentation methods robust to heterogeneous imaging and efficient for routine deployment [3]. While fundamental to 3D quantitative imaging tasks like organ volumetry and therapy planning [15,18,8], challenges from domain shift and the high computational footprint of modern 3D backbones hinder resource-efficient deployment [3,4,13,24].

Knowledge distillation (KD) provides a principled route to mitigate the computational burden of 3D models by transferring supervision from high-performing

[*] Equal contribution.
[†] Corresponding authors.

teachers to efficient student segmenters [7,22,20,14]. However, KD for dense 3D segmentation faces distinctive obstacles. First, voxel-wise distillation (e.g., logit or feature alignment) assumes that teacher and student representations are structurally compatible. Under architectural heterogeneity, differing feature scales or semantics can render strict activation matching unstable [16,5,25]. Second, 3D segmentation is often severely class-imbalanced: most voxels belong to background, so uniform alignment or naive pooling can dilute supervision, especially for small targets [26]. Third, medical imaging data are subject to cross-scanner appearance drift. In this setting, absolute activation imitation may overfit to style-specific cues and acquisition idiosyncrasies rather than preserving transferable, clinically meaningful structures [3,2,10].

These limitations suggest that effective distillation for 3D segmentation should rely less on absolute, voxel-aligned activation imitation and more on transferable structure. Beyond producing accurate per-case outputs, a strong teacher often induces a representation space where clinically similar cases form coherent neighborhoods [28,30]. Prior work has shown that distilling inter-sample relations (e.g., pairwise similarities or distances) can provide a softer, architecture-agnostic signal than pointwise feature matching [21,28,27,16,1,12]. However, applying existing relational KD methods from natural imaging to 3D medical data presents significant bottlenecks. Conventional RKD [21] transfers structural knowledge by penalizing differences in scalar distances or angles, which collapses high-dimensional anatomical variations into isotropic scalars. Similarly, CIRKD [29] constructs dense pixel-level relations for 2D images; however, this becomes computationally prohibitive and inherently noisy in 3D segmentation, where volumes are overwhelmingly dominated by non-informative background voxels.

To address these gaps, we propose *Displacement Preserving Relational Distillation (DPRD)* for robust 3D medical segmentation. DPRD performs vector displacement matching on ROI-masked case-level embeddings across hierarchical encoder stages. By pooling encoder features into task-relevant ROI embeddings and aligning all-pairs batch-wise displacements, DPRD avoids direct activation matching and scalar-only relational supervision while reducing background dilution. Coupling trajectory alignment with batch-wise scale normalization provides scale-invariant local relational supervision. Ablations support the contributions of trajectory modeling and ROI masking. Our contributions are:

- We introduce *displacement-preserving* relational distillation for 3D segmentation, using batch-wise displacement vectors with relational scale normalization to impose a scale-invariant local relational constraint rather than matching absolute activations.
- We develop a practical cross architecture, ROI-aware, multi-stage distillation design that is memory efficient for 3D volumes and explicitly mitigates background dominated supervision.
- We empirically show consistent improvements across ISLES and AMOS benchmarks and provide ablation analysis on the contributions of ROI-aware masking, displacement alignment, and distance consistency.

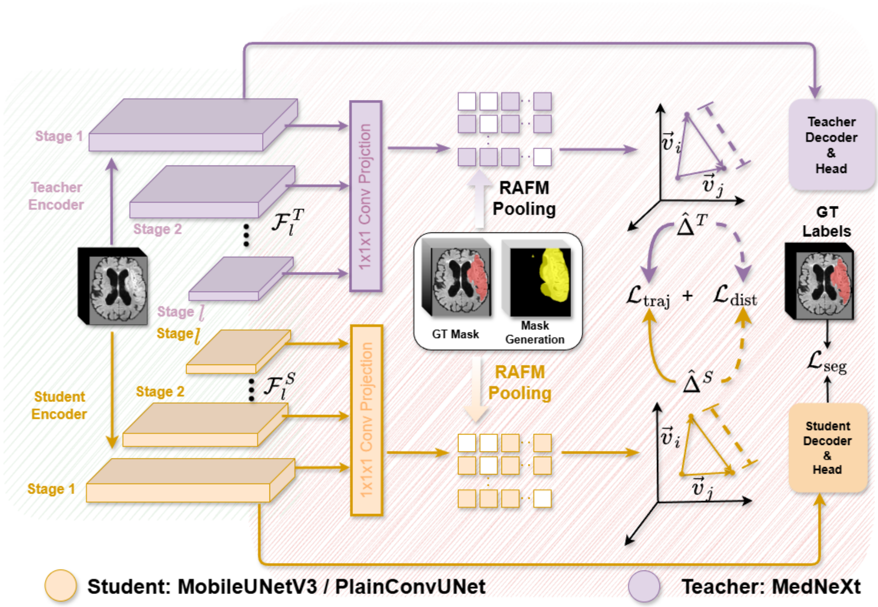

Fig. 1. Overview of DPRD. RAFM obtains ROI-aware case-level embeddings, and DPRA aligns pairwise displacement relations between teacher and student representations.

2 Method

We propose Displacement-Preserving Relational Distillation (DPRD) for high performance, lightweight 3D medical segmentation (Fig. 1). Unlike restrictive voxel-wise alignment, DPRD matches batch-wise displacement vectors after relational scale normalization in a common projected embedding space, providing scale-invariant local relational supervision between teacher and student representations. To handle architectural and scanner variability, we introduce: (1) *Displacement Preserving Relational Alignment*, which matches teacher-induced local displacement relations instead of absolute activations; and (2) *ROI-Aware Feature Masking*, which uses anatomical priors to focus alignment on task-relevant structures and mitigate background dilution. These modules ensure multi-stage structural consistency, enabling the student to learn 3D anatomical variances.

2.1 Displacement-Preserving Relational Alignment (DPRA)

Pairwise Displacement Mapping To facilitate architecture-agnostic knowledge transfer, we first extract a set of feature embeddings $\{\mathbf{v}_i^\phi\}_{i=1}^N$ $(\phi \in \{T, S\})$ (Detail in Sec 2.2) from the l-th stage of the teacher and student encoders. Unlike conventional relational knowledge distillation that collapses complex sample relationships into scalar distances or angles, our method models local relational cues by calculating pairwise displacement vectors between ROI-pooled case-level embeddings for all unique pairs (i, j):

$$\Delta_{ij}^{\phi} = \mathbf{v}_i^{\phi} - \mathbf{v}_j^{\phi}, \quad \phi \in \{T, S\} \tag{1}$$

where $1 \leq i < j \leq N$, for a mini-batch of size N, this yields $K = N(N-1)/2$ displacement vectors. In the teacher's latent space, Δ_{ij}^{T} represents a high-dimensional relational bridge that captures structural patterns less dependent on specific convolutional operations. The student's task is to reconstruct these "anatomical trajectories" within its own latent space, ensuring that the directional evolution from one case to another remains consistent across different feature hierarchies.

Heterogeneous Batch Normalization To ensure stable convergence when distilling knowledge between architectures with different activation scales, we introduce a batch-wise normalization scheme to eliminate intrinsic activation bias. For each stream ϕ, we calculate the batch-level relational scale μ^{ϕ} as the average L_2 norm across all K displacement vectors within the current batch:

$$\mu^{\phi} = \frac{1}{K} \sum_{1 \leq i < j \leq N} \|\Delta_{ij}^{\phi}\|_2 \tag{2}$$

The normalized displacement vectors are then defined as:

$$\hat{\Delta}_{ij}^{\phi} = \Delta_{ij}^{\phi} / (\mu^{\phi} + \epsilon) \tag{3}$$

This normalization removes global activation-scale differences, encouraging the student to focus on scale-invariant local displacement relations rather than absolute feature magnitudes. By aligning the scaled displacements, we encourage the distillation process to exhibit enhanced stability against intensity fluctuations and scale discrepancies common in multi-vendor medical imaging data.

Relational Distillation Objective To provide scale-invariant local relational supervision, the relational loss $\mathcal{L}_{rel}^{l}$ at stage l is composed of two synergistic components:

(1) **Pairwise Displacement Loss ($\mathcal{L}_{traj}$):** This term functions as a fine-grained trajectory reconstruction, constraining the student to align the precise orientation and normalized length distribution of the teacher's displacement vectors across all latent dimensions. We compute the cumulative error across the shared channel dimension C using the outlier-robust Smooth L1 loss Ψ:

$$\mathcal{L}_{traj} = \frac{1}{K} \sum_{1 \leq i < j \leq N} \left(\sum_{c=1}^{C} \Psi(\hat{\Delta}_{ij,c}^{S} - \hat{\Delta}_{ij,c}^{T}) \right) \tag{4}$$

(2) **Distance Consistency Regularization ($\mathcal{L}_{dist}$):** While the aforementioned normalization reduces overall scaling differences, $\mathcal{L}_{dist}$ explicitly preserves the relative length structure within the batch by aligning the relative Euclidean distances of the displacements. This provides a complementary constraint on the relational scale:

$$\mathcal{L}_{dist} = \frac{1}{K} \sum_{1 \le i < j \le N} \Psi(\|\hat{\Delta}_{ij}^S\|_2 - \|\hat{\Delta}_{ij}^T\|_2) \tag{5}$$

The total stage-wise relational loss is defined as $\mathcal{L}_{rel}^l = \lambda_{traj}\mathcal{L}_{traj} + \lambda_{dist}\mathcal{L}_{dist}$, where λ_{traj} and λ_{dist} are balancing hyperparameters. This composite objective encourages the student to match teacher-induced local displacement directions and relative distance patterns without requiring global manifold preservation.

2.2 ROI-Aware Feature Masking (RAFM)

Density-Normalized Feature Aggregation To bridge the architectural discrepancy between the teacher and student, we project their feature maps into a shared 128-channel space via $1\times1\times1$ convolutions, yielding channel-aligned maps $\hat{\mathbf{F}}^T$ and $\hat{\mathbf{F}}^S$. We then employ a density-normalization scheme to produce size-invariant embeddings from these projected maps. This ensures that the subsequent displacement vectors represent inter-case anatomical variance rather than differences in voxel count or occupied volume.

$$\mathbf{v}^\phi = \frac{\mathcal{P}(\hat{\mathbf{F}}^\phi \odot \mathbf{M}^\phi)}{\mathcal{P}(\mathbf{M}^\phi) + \epsilon} \tag{6}$$

where $\mathcal{P}(\cdot)$ denotes the pooling operation used to obtain compact case-level embeddings, and ϵ is a small constant for numerical stability. By reweighting the pooled features by the mask density $\mathcal{P}(\mathbf{M}^\phi)$, this strategy yields a compact, scale-invariant representation that encapsulates the average activation intensity within the anatomical structures, providing a stable foundation for relational alignment.

Anatomical Mask Construction To ground distillation in relevant anatomy, we derive a unified binary mask $\mathbf{M}$ by merging all foreground ground-truth categories:

$$\mathbf{M}_{raw} = \mathbb{1}\left(\sum_{c=1}^{C_{class}} \mathbf{Y}_c > 0\right) \tag{7}$$

where $\mathbf{Y}$ represents the segmentation target. To align with the respective spatial resolutions of the teacher and student feature maps, $\mathbf{M}_{raw}$ is rescaled using nearest-neighbor interpolation to generate model specific masks: $\mathbf{M}^T$ and $\mathbf{M}^S$. This localization ensures that the anatomical ROI is precisely captured within each model's unique coordinate system before the aggregation process. It is important to note that the ROI-aware masking is exclusively utilized during the training phase to guide the relational alignment.

2.3 Multi-Stage Joint Optimization

To capture hierarchical anatomical structures, we apply DPRD across all stages in the encoder set L. We employ stage-specific coefficients to balance the relational signals across varying feature resolutions. The student model is optimized by minimizing a composite loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \sum_{l \in L} \omega^l \left(\lambda_{traj} \mathcal{L}_{traj}^l + \lambda_{dist} \mathcal{L}_{dist}^l \right) \tag{8}$$

where $\mathcal{L}_{seg}$ denotes the standard nnU-Net objective (Dice and Cross-Entropy). Stage-wise weights ω^l for $l \in \{0, \ldots, 5\}$ are set to $\{0.05, 0.1, 0.1, 0.15, 0.25, 0.35\}$. We assign higher weights to deeper layers, which typically encode more abstract anatomical patterns. Balancing weights are fixed at $\lambda_{traj} = 1.0$ and $\lambda_{dist} = 0.5$. This multi-scale consistency ensures the student prioritizes anatomical variances robust to acquisition noise and heterogeneous multi-center benchmarks.

3 Experiment and Results

Datasets and Metrics
We evaluate DPRD on ISLES 2022 [6] (200 training/50 validation cases for MRI stroke lesion segmentation) and AMOS 2022 [9] (240 training/60 validation cases for 15-class abdominal organ segmentation). Performance is assessed via Dice Similarity Coefficient (DSC), 95% Hausdorff Distance (HD95), and Normalized Surface Distance (NSD).
Implementation Details.
All experiments are implemented in nnU-Net [8]. For cross-architecture distillation, we use a pre-trained, frozen MedNeXt as the teacher and evaluate two heterogeneous students: MobileUNetV3 and PlainConvUNet. Following the nnU-Net validation protocol, results are reported on the standard data partitioning. Batch size is set to **4** for MobileUNetV3 and **2** for PlainConvUNet due to heterogeneous 3D memory footprints. DPRD provides stochastic local relational supervision, where mini-batch displacement constraints are accumulated across sampled cases, augmented patches, encoder stages, iterations, and epochs. The $O(N^2)$ relational alignment is lightweight because it operates on compact case-level embeddings rather than dense voxel-wise feature maps. During inference, the teacher and projection layers are removed, so the deployed model has the same parameters and FLOPs as the corresponding student. All models are trained with SGD. We compare DPRD with Logits KD [7], FitNet [23], RKD [21], and CIRKD [29], implemented under the same nnU-Net training setting with hyperparameters adapted for 3D medical volumes.

3.1 Results

Quantitative Results Tables 1–2 summarize comparisons on ISLES 2022 and AMOS (CT) against standard distillation baselines. On ISLES 2022, DPRD

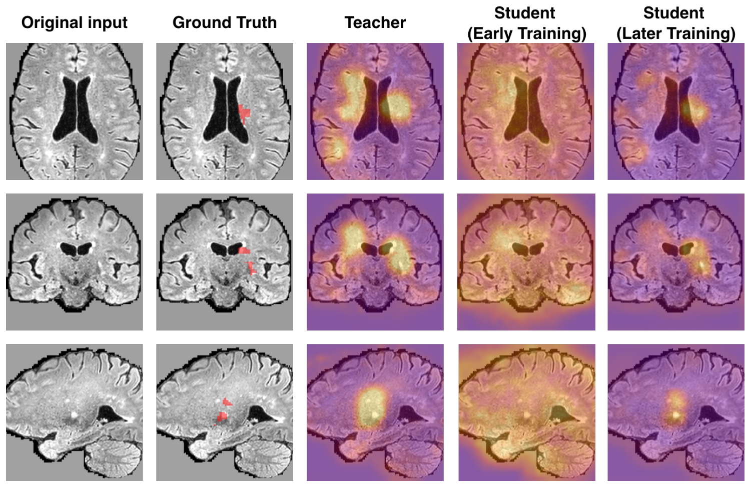

Fig. 2. Visualization of DPRD on ISLES 2022 at encoder stage 3 for a sample case in each orthogonal view (axial, coronal, sagittal). Columns: input MRI, ground truth (red), teacher, early-stage student, and late-stage student responses. Yellowish regions denote areas of higher importance; the late-stage student focuses on lesion regions more strongly than the early-stage student and aligns better with the ground truth.

achieves the best Dice (75.16%) and improves the PlainConvUNet student baseline (72.92% Dice, 17.21 mm HD95) with the best surface scores among all methods (85.92% NSD, 13.08 mm HD95), the latter even surpassing the MedNeXt teacher (14.12 mm HD95). On AMOS, DPRD delivers the best overall performance (85.46% Dice, 85.29% NSD, 11.72 mm HD95), outperforming all distillation baselines and slightly exceeding the MedNeXt teacher on Dice (85.37%) while reducing boundary error (HD95 14.52→11.72 mm). Overall, DPRD consistently strengthens compact students, with particularly pronounced gains on boundary-sensitive metrics.

Qualitative Results Fig. 2 visualizes encoder-stage responses on ISLES, where the late-stage student shows stronger focus on lesion regions than the early-stage student and better agreement with the ground-truth annotations. Fig. 3 further compares AMOS segmentations across KD baselines; DPRD produces masks more consistent with the ground truth, particularly around the liver, postcava, and gallbladder boundaries.

Ablation Study Table 3 shows the contribution of each DPRD component. Removing RAFM decreases Dice from 85.46% to 81.68% and increases HD95 to 21.65 mm, indicating that background-dominated features can dilute the relational distillation signal. Removing either trajectory mapping or distance regularization also degrades performance, with HD95 increasing to 15.03 mm and 17.07 mm, respectively. The full *DPRD* framework achieves the best HD95 of 11.72 mm, suggesting that directional trajectory reconstruction and relative dis-

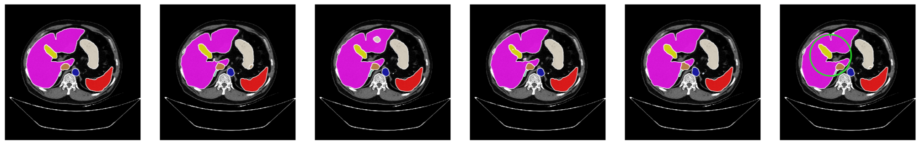

Fig. 3. Qualitative comparison on AMOS. Colors indicate different organs and are consistent across columns. Compared to other distillation baselines, our **DPRD** produces segmentation masks that are more consistent with the ground truth, particularly for the structures highlighted by the green circles (e.g., gallbladder and postcava).

Table 1. Comparison of distillation methods on ISLES 2022.

Method	Params(M)	FLOPs(G)	Dice(%) ↑	NSD(2mm) ↑	HD95(mm) ↓
Teacher-MedNeXt	12.10	974.58	76.88±2.84	85.64±2.55	14.12±3.11
Baseline-PlainConvUNet	1.95	497.79	72.92±3.09	83.34±2.87	17.21±3.57
Logits KD [7]			72.02±3.22	82.61±3.12	17.49±3.74
FitNet [23]			73.93±3.05	84.29±2.82	16.79±3.57
RKD [21]	1.95	497.79	72.85±3.15	83.51±2.94	18.42±4.03
CIRKD [29]			74.04±3.03	85.06±2.92	14.43±3.53
Ours(DPRD)			**75.16±3.03**	**85.92±2.90**	**13.08±3.37**

Table 2. Comparison of distillation methods on AMOS (CT) 2022.

Method	Params(M)	FLOPs(G)	Dice(%) ↑	NSD(2mm) ↑	HD95(mm) ↓
Teacher-MedNeXt	12.11	1360	85.37±2.03	84.93±2.14	14.52±2.12
Baseline-MobileUNetV3	0.63	39.19	80.16±2.58	77.87±3.09	18.74±1.69
Logits KD [7]			83.14±2.42	81.80±2.74	15.60±1.60
FitNet [23]			81.80±2.74	80.01±2.68	15.02±1.89
RKD [21]	0.63	39.19	83.49±2.29	82.77±2.42	17.15±2.43
CIRKD [29]			82.09±2.54	80.13±2.67	14.75±1.46
Ours(DPRD)			**85.46±2.13**	**85.29±2.14**	**11.72±2.11**

tance consistency provide complementary supervision for accurate anatomical surface segmentation.

4 Conclusion

In this paper, we propose DPRD, an efficient distillation framework for robust 3D medical image segmentation. By combining ROI-aware feature masking with displacement-preserving relational alignment, DPRD improves compact students over feature-based and relational baselines on ISLES and AMOS, while achieving performance comparable to the MedNeXt teacher with substantially lower inference cost. A current limitation is that RAFM relies on ground-

Table 3. DPRD ablation study on AMOS 2022 (MobileUNetV3)

Method	RAFM	Traj.	Dist.	Dice (%) ↑	NSD (2mm) ↑	HD95 (mm) ↓
Baseline	–	–	–	80.16±2.58	77.87±3.09	18.74±1.69
RAFM-off	–	✓	✓	81.68±2.67	79.72±2.84	21.65±2.74
Distance-off	✓	✓	–	82.98±2.45	82.02±2.48	17.07±2.23
Trajectory-off	✓	–	✓	83.04±2.39	82.21±2.42	15.03±2.03
DPRD	✓	✓	✓	**85.46±2.13**	**85.29±2.14**	**11.72±2.11**

truth label masks during training, which may restrict its applicability in weakly-supervised or label-scarce settings; future work will explore teacher-predicted ROIs or uncertainty-guided masks to relax this requirement.

Acknowledgments. The authors gratefully acknowledge the computational resources provided by the Delta system at the National Center for Supercomputing Applications (NCSA) [award OAC 2005572] through ACCESS allocations CIS260331 and CIS250976. ACCESS, the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support program, is supported by U.S. National Science Foundation (NSF) grants #2138259, #2138286, #2138307, #2137603, and #2138296. This work was partially supported by NIH grant U01AG079847 and NSF Award No. 2412285.

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

References

1. Ding, Z., Chu, X., Tian, Q.: Difference feature map distillation: Transferring inter-sample relational knowledge towards efficient transformer-based tracking. In: Proceedings of the 2026 IEEE Intelligent Vehicles Symposium (IV) (2026)
2. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging **39**(7), 2415–2425 (2020)
3. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
4. 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)
5. He, T., Shen, C., Tian, Z., Gong, D., Sun, C., Yan, Y.: Knowledge adaptation for efficient semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 578–587 (2019)
6. Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific data **9**(1), 762 (2022)
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

8. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
9. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
10. Jia, J., Wu, Y., Chen, H., Jing, H., Wang, H., Bu, J., Wu, L.: Unsupervised causal prototypical networks for de-biased interpretable dermoscopy diagnosis. arXiv preprint arXiv:2602.23752 (2026)
11. Lan, Q., Choi, A., Ma, J., Wang, B., Zhao, Z., Jiang, X., Hsu, Y.C.: From performance to practice: Knowledge-distilled segmentator for on-premises clinical workflows. arXiv preprint arXiv:2601.09191 (2026)
12. Lan, Q., Tian, Q.: Acam-kd: adaptive and cooperative attention masking for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3957–3966 (2025)
13. 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)
14. Li, Y., Meng, S., Yang, C., Feng, W., Liu, J., An, Z., Wang, Y., Tian, Y.: A comprehensive survey of interaction techniques in 3d scene generation. Authorea Preprints (2026)
15. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
16. Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J.: Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2604–2613 (2019)
17. Mezger, U., Jendrewski, C., Bartels, M.: Navigation in surgery. Langenbeck's archives of surgery **398**(4), 501–514 (2013)
18. 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). pp. 565–571. Ieee (2016)
19. Nitsch, J., Klein, J., Dammann, P., Wrede, K., Gembruch, O., Moltz, J., Meine, H., Sure, U., Kikinis, R., Miller, D.: Automatic and efficient mri-us segmentations for improving intraoperative image fusion in image-guided neurosurgery. NeuroImage: Clinical **22**, 101766 (2019)
20. Noothout, J.M., Lessmann, N., Van Eede, M.C., Van Harten, L.D., Sogancioglu, E., Heslinga, F.G., Veta, M., Van Ginneken, B., Išgum, I.: Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation. Journal of Medical Imaging **9**(5), 052407–052407 (2022)
21. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3967–3976 (2019)
22. Qin, D., Bu, J.J., Liu, Z., Shen, X., Zhou, S., Gu, J.J., Wang, Z.H., Wu, L., Dai, H.F.: Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging **40**(12), 3820–3831 (2021)
23. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015), `https://arxiv.org/abs/1412.6550`

24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 405–415. Springer (2023)
25. Shu, C., Liu, Y., Gao, J., Yan, Z., Shen, C.: Channel-wise knowledge distillation for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5311–5320 (2021)
26. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 240–248. Springer (2017)
27. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
28. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019)
29. Yang, C., Zhou, H., An, Z., Jiang, X., Xu, Y., Zhang, Q.: Cross-image relational knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12319–12328 (2022)
30. Zhu, C., Lin, Y., Shao, J., Lin, J., Wang, Y.: Pathology-aware prototype evolution via llm-driven semantic disambiguation for multicenter diabetic retinopathy diagnosis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9196–9205 (2025)