This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Detail Consistent Stage-Wise Distillation for Efficient 3D MRI Segmentation

Mengchen Fan[1], Baocheng Geng[1], Xi Xiao[1], Tianyang Wang[1], Siyuan Mei[2], Pulin Che[1], Xiaoqian Jiang[3⋆], and Qizhen Lan[3⋆]

[1] University of Alabama at Birmingham, Birmingham, AL, USA
{fanm,bgeng,xxiao,tw2,pche}@uab.edu
[2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
siyuan.mei@fau.de
[3] UTHealth Houston, Houston, TX, USA
{xiaoqian.jiang,qizhen.lan}@uth.tmc.edu

Abstract. Deploying high-performing 3D medical image segmenters (e.g., nnU-Net) is often limited by memory footprint and inference latency. Compression is therefore necessary, but compact 3D encoders tend to lose fine structural cues (small lesions and sharp boundaries) as downsampling repeats across multi-resolution stages. We propose *Detail Consistent Distillation (DCD)*, a stage-wise distillation framework that preserves structural detail across scales by aligning teacher–student features in a wavelet-decomposed representation. At each encoder stage, DCD distills directional detail components in the wavelet domain while leaving the coarse approximation comparatively unconstrained, avoiding over-regularization of global semantics. DCD is used only during training and introduces no inference-time overhead. Experiments on the BraTS 2024 and ISLES 2022 benchmarks demonstrate that our approach achieves superior performance in MRI segmentation using 3D multi-modal data. Code and implementation details for DCD are publicly available at `https://github.com/ClinicaAlpha/DCD-3D-MedSeg`.

Keywords: knowledge distillation · 3D medical segmentation

1 Introduction

Accurate 3D lesion segmentation in brain MRI supports quantitative neuroimaging and time-sensitive workflows, including tumor sub-region assessment and ischemic stroke lesion delineation [17,10]. Encoder–decoder U-Net designs [22,3] and standardized training pipelines such as nnU-Net [12] have made strong 3D baselines widely accessible, but volumetric models remain expensive in memory and latency, limiting deployment on constrained hardware and in high-throughput settings [27], where hardware-aware model design has become increasingly important [2,6,8,7].

⋆ Corresponding author.

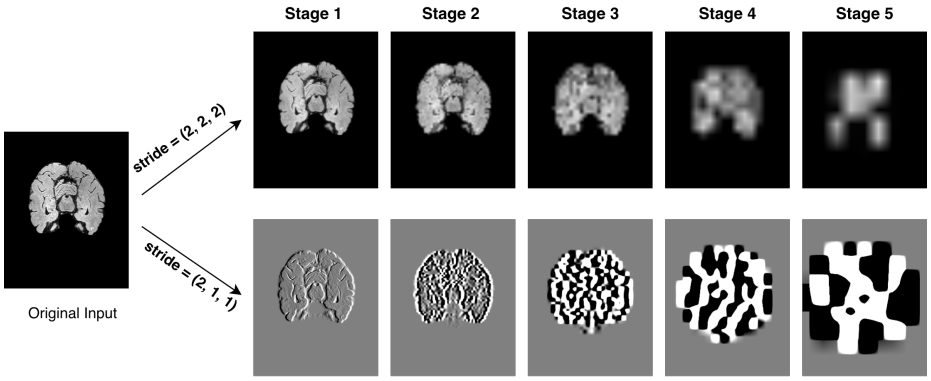

Fig. 1. Five successive downsampling stages (stride=2 in D, H, W) shown for two complementary responses. The intensity-like stream (top row) progressively loses fine anatomical details, while the derivative/edge response (bottom row, D_x) rapidly degrades into coarse, aliased patterns, illustrating joint loss of global structure and local detail under repeated decimation.

Model compression is therefore necessary. Knowledge Distillation (KD) transfers behavior from a strong teacher to a compact student during training [1,11] and has been used to compress segmentation ensembles into smaller networks while retaining accuracy [15,14]. However, compact 3D segmenters often fail in a characteristic way: they preserve coarse localization but lose fine boundaries and small or thin structures. We attribute this to progressive suppression of detail-sensitive representations as capacity shrinks and downsampling repeats over the encoder hierarchy.

Two effects motivate explicit supervision of detail components in compressed 3D UNets. First, strided downsampling is a lossy sampling operation (as shown in Fig. 1). It can discard or alias high-frequency content with errors compounding across stages [33]. Second, deep networks exhibit spectral bias: they fit lower-frequency components earlier and more readily, while higher-frequency variation is harder to represent and typically requires more capacity [20,31]. Together, these effects make conventional "match everything" distillation prone to producing over-smoothed students even when the teacher's predictions are sharp.

Frequency-domain supervision provides a direct handle on the components most vulnerable to over-smoothing. The discrete wavelet transform (DWT) decomposes a volume into a low-frequency approximation and directional detail subbands with exact reconstruction, and wavelets are established tools for separating scale-specific information in biomedical signals and images. [16,4,25]. In MRI, this perspective is physically grounded by k-space: low spatial frequencies dominate global contrast, while higher spatial frequencies contribute disproportionately to edges and fine detail [19,18]. However, not all high frequency content is equally informative; high frequency bands can also capture acquisition noise and scanner-specific artifacts, and MRI statistics differ from natural scenes [29].

This motivates selective frequency supervision that emphasizes boundary relevant detail while avoiding the noise dominated components.

Recent KD work has explored frequency-aware distillation in 2D natural image settings [13,32,34], but medical images have different spectral statistics than natural scenes [9,23,29], so indiscriminate frequency matching can align noise or acquisition-specific artifacts rather than clinically meaningful structure.

Motivated by this, we propose **Detail Consistent Stage-Wise Distillation (DCD)** for efficient 3D medical image segmentation. DCD aligns teacher and student encoder features at each resolution stage by applying a 3D DWT to isolate wavelet subbands and distilling only directional detail components, while keeping the low frequency approximation comparatively unconstrained to preserve global semantics. To maintain spatially interpretable supervision, we reconstruct detail-only features via inverse DWT (IDWT) before performing alignment, yielding a spatial-domain loss that targets detail content. DCD is used only during training and introduces no inference-time overhead. We evaluate on brain tumor and stroke lesion segmentation benchmarks [17,10] and report improved accuracy–efficiency trade-offs under nnU-Net compression. Our main contributions are:

1. A wavelet-domain subband selection strategy that distills directional detail while excluding the most noise-prone extreme high-frequency band;
2. A stage-wise distillation framework that explicitly targets detail subspaces in compressed 3D U-Net architectures; and
3. Empirical evaluation on 3D brain tumor and stroke lesion benchmarks showing improved boundary-sensitive quality under compression, without adding inference-time cost.

2 Proposed Method

We distill a compact student model from a high-capacity teacher (as shown in Fig. 2) by explicitly emphasizing detail bearing responses that tend to degrade under compression and repeated downsampling. Given an input volume x, let $F_i^{(t)} \in \mathbb{R}^{C_i^{(t)} \times H_i \times W_i \times D_i}$ and $F_i^{(s)} \in \mathbb{R}^{C_i^{(s)} \times H_i \times W_i \times D_i}$ denote encoder stage features of the teacher and student at stage $i \in \{1, \ldots, M\}$. The student is obtained by uniform channel reduction, i.e., $C^{(s)} = C^{(t)}/r$ for a reduction factor r. To align channel dimensions for distillation, we use a learnable $1 \times 1 \times 1$ projection $\phi_i(\cdot)$. The teacher parameters are frozen during distillation.

2.1 Capacity-Induced Spectral Bias

When model capacity is reduced, the student tends to preserve components that dominate the loss and feature energy, while fine structural variations are more easily suppressed. This behavior is consistent with the spectral bias of neural networks. During gradient based training, lower frequency components are often learned more readily than higher frequency variation, which can require greater representational capacity [20,31,30,5].

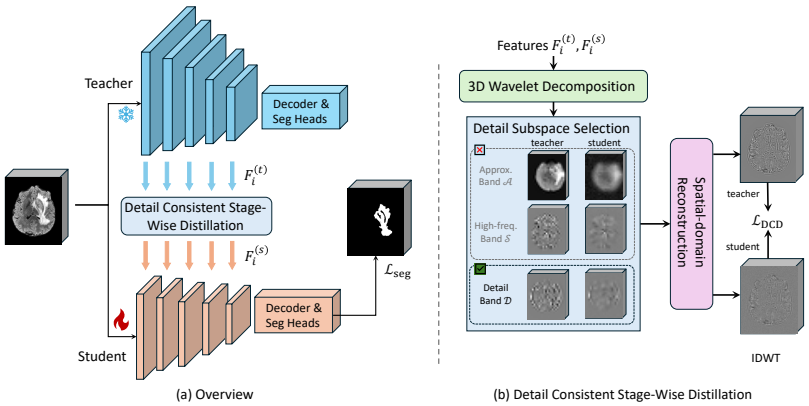

Fig. 2. Overview of Detail Consistent Distillation (DCD), (a) Teacher–student distillation in a U-Net-like architecture, where DCD is applied at each encoder stage and optimized jointly with the segmentation loss. (b) Stage-wise DCD. Given stage features, we perform a 3D wavelet decomposition, select the detail subspace (band $\mathcal{D}$), reconstruct detail only features via IDWT, and compute $\mathcal{L}_{\mathrm{DCD}}$ as the MSE between teacher and student.

Therefore, KD in U-Net-like architectures inherently exhibits a spectral bias. In U-Net-like architectures that repeatedly downsample along the encoder hierarchy, this bias can manifest as students that match coarse semantics while under representing boundary critical detail.

We summarize the above mismatch as the following motivating inequality between teacher and student features at stage i:

$$\left\|\mathcal{P}_{\mathrm{hi}}\left(F_i^{(t)}\right)-\mathcal{P}_{\mathrm{hi}}\left(F_i^{(s)}\right)\right\|_2 \;>\; \left\|\mathcal{P}_{\mathrm{lo}}\left(F_i^{(t)}\right)-\mathcal{P}_{\mathrm{lo}}\left(F_i^{(s)}\right)\right\|_2, \tag{1}$$

where $\mathcal{P}_{\mathrm{hi}}(\cdot)$ and $\mathcal{P}_{\mathrm{lo}}(\cdot)$ denote complementary projections onto high- and low-frequency components.

2.2 Wavelet Detail Subspace Selection

Wavelet Subspace Decomposition To obtain a structured multi-resolution decomposition, we apply a 3D Discrete Wavelet Transform (DWT) independently to each channel of a stage feature tensor $F \in \mathbb{R}^{C\times H\times W\times D}$:

$$\left\{F_l^{(\beta)}\right\}_{\beta\in\{L,H\}^3} = \xi_l(F), \qquad \beta \in \{LLL, LLH, \ldots, HHH\}, \tag{2}$$

We use an l-level decomposition by recursively applying the transform to the approximation subband LLL, $\xi_l(\cdot)$ is the level l 3D-DWT and β indexes low-/high-pass filtering along the three spatial axes. In this paper, we set $l = 3$ for all distillation experiments.

For distillation, we partition the subbands as

$$\mathcal{A} = \{LLL\}, \qquad \mathcal{S} = \{HHH\}, \qquad \mathcal{D} = \{L, H\}^3 \setminus \{LLL, HHH\}. \tag{3}$$

$\mathcal{A}$ captures coarse structure, while $\mathcal{D}$ contains detail components that are strongly coupled to boundaries and thin structures, and $\mathcal{S}$ represents simultaneously high-pass filtering along all three axes and is often more noise-sensitive in MRI.

Proposition 1 (Motivation for excluding $\mathcal{S}$ under a simplified MRI noise model). *Assume a stage feature can be decomposed as $F = F_{\text{sig}} + \epsilon$, where ϵ is additive, zero-mean, i.i.d. noise with variance σ^2 per voxel (a standard white-noise approximation for high-frequency noise components). $\xi(\cdot)$ is orthonormal. Then energy is preserved and the expected noise energy per subband satisfies*

$$\|\xi(\epsilon)\|_2^2 = \|\epsilon\|_2^2, \qquad \mathbb{E}\Big[\|\xi(\epsilon)^{(\beta)}\|_2^2\Big] = \sigma^2 N_\beta, \tag{4}$$

where N_β is the number of coefficients in subband β. In MRI, signal energy often decays with spatial frequency (k-space center dominance), so the expected signal energy in the band $\mathcal{S} = \{\text{HHH}\}$ can be substantially smaller than that in the directional detail bands $\beta \in \mathcal{D}$,

$$\mathbb{E}\Big[\|\xi(F_{\text{sig}})^{(\mathcal{S})}\|_2^2\Big] \ll \max_{\beta \in \mathcal{D}} \mathbb{E}\Big[\|\xi(F_{\text{sig}})^{(\beta)}\|_2^2\Big]. \tag{5}$$

Defining the subband SNR as $\text{SNR}_\beta = \frac{\mathbb{E}\left[\|\xi(F_{\text{sig}})^{(\beta)}\|_2^2\right]}{\sigma^2 N_\beta}$, which implies $\text{SNR}_{\mathcal{S}}$ is typically much smaller than SNR_β for $\beta \in \mathcal{D}$ under the condition in Eq. 5. This motivates excluding $\mathcal{S}$ from distillation to avoid encouraging the student to learn noise dominated components.

Subband $\mathcal{D}$ selection Full spectrum feature matching is typically dominated by high energy low-frequency components, which weakens supervision on detail-bearing bands and amplifies the capacity-induced mismatch in Eq. (1). DCD therefore removes the distillation constraint on the approximation band $\mathcal{A}$, but we do not remove low frequency features from the network, so the student is not over-regularized on global semantics. Moreover, by Proposition 1, the band $\mathcal{S}$ tends to have lower effective SNR and can inject noise into the distillation signal. Therefore, we choose $\mathcal{D}$ as the distillation frequency subspace to preserve directional details while avoiding both low-frequency dominance $\mathcal{A}$ and noise-prone components $\mathcal{S}$.

2.3 Detail Consistent Stage-Wise Distillation

We define a detail projection that reconstructs a spatial-domain representation using only the detail subbands $\mathcal{D}$:

$$\mathcal{P}_{\mathcal{D}}(F) = \xi_l^{-1}\Big(\Big\{\widetilde{F}_l^{(\beta)}\Big\}\Big), \qquad \widetilde{F}_l^{(\beta)} = \begin{cases} F_l^{(\beta)}, & \beta \in \mathcal{D}, \\ 0, & \beta \notin \mathcal{D}, \end{cases} \tag{6}$$

where $\xi_l^{-1}(\cdot)$ denotes the 3D IDWT. This construction restricts supervision to detail content while producing a spatial-domain tensor, so the alignment loss is computed in the same geometry as standard feature MSE (rather than directly on coefficient tensors whose layout depends on implementation choices).

At each stage i, we align teacher and student projections as

$$\mathcal{L}_{\mathrm{DCD}}^{(i)} = \frac{1}{N_i} \left\| \mathcal{P}_{\mathcal{D}}\left(F_i^{(t)}\right) - \mathcal{P}_{\mathcal{D}}\left(\phi_i\left(F_i^{(s)}\right)\right) \right\|_2^2, \tag{7}$$

The multi-stage distillation loss sums across stages: $\mathcal{L}_{\mathrm{DCD}} = \sum_{i=1}^{M} \mathcal{L}_{\mathrm{DCD}}^{(i)}$. The overall training objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \mu\, \mathcal{L}_{\mathrm{DCD}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{seg}}$ is the segmentation objective (Dice + cross-entropy) and μ balances distillation. Since DWT/IDWT operators and ϕ_i are used only during training, DCD adds no inference-time overhead. Fig. 2(b) shows the overall workflow of single-stage DCD distillation.

3 Experiment

3.1 Dataset and Implementation

Datasets We utilized the BraTS2024-BraTS-GLI dataset [26] and the ISLES 2022 dataset [10]. The BraTS 2024 dataset contains 1,350 cases (1,080 for training and 270 for validation). Each case includes four MRI modalities: pre-contrast T1-weighted (T1n), contrast-enhanced T1-weighted (T1c), T2-weighted (T2w), and T2-FLAIR (T2f). Tumor annotations are divided into four subregions: enhancing tissue (ET), non-enhancing tumor core (NETC), surrounding non enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC). The ISLES 2022 dataset contains 250 cases (200 for training and 50 for validation), and each case includes three MRI modalities: apparent diffusion coefficient (ADC), diffusion-weighted imaging (DWI), and FLAIR. Its annotations consist of a single category: lesion.

Implementation details. The proposed framework is implemented in PyTorch and trained on a single NVIDIA A100 GPU. We adopt SGD with Nesterov momentum ($m = 0.99$) using an initial learning rate of 0.01 and weight decay of 3×10^{-5}. We build the compact student using a channel reduction factor $r = 4$. Both teacher and student follow the nnU-Net encoder and decoder topology. For Detail Consistent distillation, we use the Daubechies-4 wavelet (db4) as the wavelet basis, distilling the directional-detail subbands $\mathcal{D}$ at all encoder stages. The distillation weight is 0.05.

3.2 Experiment Results

Comparative Results. All results reported in the tables are presented as mean ± standard error of the mean. Table 1 demonstrates that the proposed DCD

Table 1. Performance comparison on BraTS 2024 and ISLES 2022.

Method	BraTS 2024 (Overall)			ISLES 2022 (Overall)		
	mDice(%)↑	HD95(mm)↓	NSD(%)↑	mDice(%)↑	HD95(mm)↓	NSD(%)↑
Teacher	75.23±0.90	5.33±0.38	87.43±0.77	76.66±2.37	9.04±1.84	83.82±2.80
w/o KD	63.60±0.93	7.84±0.46	76.78±0.83	70.21±3.03	18.94±3.82	81.22±2.65
CWD [24]	63.13±0.94	8.29±0.52	75.54±0.86	71.94±2.95	9.48±2.02	82.74±2.67
IFVD [28]	66.54±0.94	7.09±0.45	79.41±0.86	72.43±2.98	11.90±2.51	83.30±2.69
Logits [11]	65.31±0.90	8.43±0.51	78.62±0.80	71.51±3.03	12.75±2.82	82.66±2.72
Feature [21]	64.39±0.96	7.59±0.48	77.10±0.87	72.70±2.90	10.45±3.14	83.58±2.67
FreeKD [34]	62.16±0.96	8.91±0.53	74.75±0.88	71.03±3.05	12.36±2.49	81.86±2.72
Ours	**68.51**±0.92	**6.25**±0.36	**81.34**±0.82	**73.95**±2.70	**12.95**±2.49	**83.47**±2.78

surpasses existing distillation baselines on both BraTS 2024 and ISLES 2022. On BraTS 2024, DCD improves the compact student from 63.60% to 68.51% mDice +4.91% (paired t-test $p = 8.64 \times 10^{-13}$, Wilcoxon signed-rank test $p = 4.40 \times 10^{-16}$). Compared with the previous best method (e.g., IFVD with 66.54% mDice), DCD improves mDice by +1.97% (paired t-test $p = 0.0078$, Wilcoxon signed-rank test $p = 0.0080$). On ISLES 2022, DCD achieves the highest 73.95% mDice, improving over the non-distilled student by +3.74% (paired t-test $p = 0.0058$, Wilcoxon signed-rank test $p = 0.0008$). Table 2 further suggests that DCD yields larger mDice gains on detail-sensitive regions. For example, NETC improves from 41.26% to 54.36% mDice +13.10%.

Table 2. BraTS 2024 mDice results on NETC/SNFH and ET/RC.

Method	NETC mDice(%)↑	SNFH mDice(%)↑	ET mDice(%) ↑	RC mDice(%)↑
Teacher	60.85±3.20	90.17±0.59	76.18±2.08	73.75±2.07
w/o KD	41.26±2.81	87.00±0.67	61.57±2.37	64.61±2.17
CWD [24]	40.97±2.77	87.26±0.69	59.91±2.38	64.39±2.17
IFVD [28]	46.34±2.90	88.23±0.64	63.95±2.34	67.67±2.14
Logits [11]	51.96±3.04	86.91±0.71	57.97±2.36	64.41±2.17
Feature [21]	42.55±2.82	87.80±0.66	60.36±2.39	66.85±2.11
FreeKD [34]	39.75±2.75	86.39±0.70	58.71±2.38	63.80±2.17
Ours	**54.36**±3.20	**88.49**±0.63	**62.55**±2.36	**68.62**±2.12

Table 3. Model complexity comparison on BraTS 2024 and ISLES 2022.

Model	BraTS 2024		ISLES 2022	
	Params (M)↓	FLOPs (TF)↓	Params (M)↓	FLOPs (TF)↓
Teacher	101.945	19.505	102.351	17.707
Student	**6.381**	**1.277**	**6.406**	**1.149**

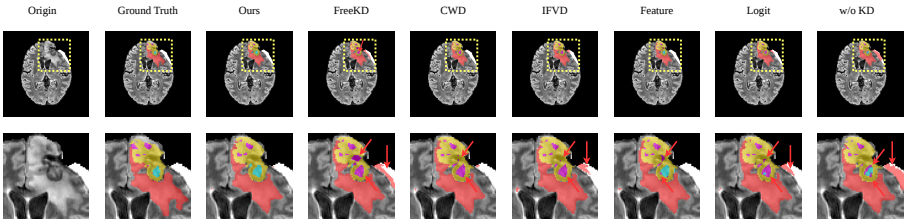

Fig. 3. Qualitative comparison on BraTS 2024 dataset. The top row shows the segmentation predictions from different methods, and the bottom row provides zoomed-in views of the regions highlighted by the yellow dashed boxes. Compared with other methods, which exhibit mis-segmentation (red arrows highlight errors), DCD produces predictions closer to the ground truth.

FreeKD, originally designed for natural images, yields limited gains for MRI segmentation, whereas DCD consistently improves mDice on both datasets. Table 3 shows that DCD uses far fewer parameters and FLOPs than the teacher while preserving competitive accuracy. Fig. 3 further shows that DCD produces predictions better aligned with the ground truth.

Table 4. Ablation study on BraTS 2024 and ISLES 2022.

Method	BraTS 2024 (Overall)			ISLES 2022 (Overall)		
	mDice(%)↑	HD95(mm)↓	NSD(%)↑	mDice(%)↑	HD95(mm)↓	NSD(%)↑
w/o KD	63.60±0.93	7.84±0.46	76.78±0.83	70.21±3.03	18.94±3.82	81.22±2.65
Band selection						
Band $\mathcal{S}$	61.98±0.96	8.91±0.53	74.80±0.90	72.12±2.96	10.40±2.12	83.26±2.62
Band $\mathcal{A}$	62.25±0.96	8.75±0.51	74.64±0.90	69.54±2.94	18.58±3.41	80.15±2.67
Band $\mathcal{D}$	**68.50±0.92**	**6.25±0.36**	**81.34±0.82**	**73.95±2.70**	**12.95±2.49**	**83.47±2.78**
IDWT						
w/o IDWT	62.87±0.91	9.01±0.51	77.13±0.84	70.06±3.11	15.69±2.79	80.51±2.80
w/ IDWT	**68.51±0.92**	**6.25±0.36**	**81.34±0.82**	**73.95±2.70**	**12.95±2.49**	**83.47±2.78**

Ablation Results. Table 4 reports ablation studies validating the contribution of each component. **(i) Subband selection.** Distilling $\mathcal{D}$ yields the best results, while distilling $\mathcal{S}$, i.e., the extreme high frequency band, performs worse than w/o KD on BraTS 2024 (61.98% vs. 63.60% mDice). This supports our motivation that the noisiest high-frequency band can provide harmful supervision. In contrast, $\mathcal{D}$ achieves 68.50% and 73.95% mDice on BraTS 2024 and ISLES 2022, outperforming both $\mathcal{S}$ and $\mathcal{A}$. **(ii) IDWT reconstruction.** Removing IDWT leads to a marked degradation on both datasets, with BraTS 2024 mDice dropping from 68.51% to 62.87% and HD95 increasing from 6.25 to 9.01, and ISLES 2022 mDice decreasing from 73.95% to 70.06%.

4 Conclusion

We propose Detail Consistent Distillation (DCD), a stage-wise distillation framework for efficient 3D medical image segmentation. DCD performs detail subspace selection at each encoder stage and distills detail subbands, enabling the student to focus on fine-grained, task-relevant cues without over-regularizing global semantics and amplifying noise. Extensive experiments on BraTS 2024 and ISLES 2022 demonstrate that DCD is robust and efficient for MRI segmentation.

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.

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

References

1. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
2. Chen, W., Wu, L., Hu, Y., Li, Z., Cheng, Z., Qian, Y., Zhu, L., Hu, Z., Liang, L., Tang, Q., Liu, Z., Yang, H.: Autoneural: Co-designing vision-language models for npu inference (2025), https://arxiv.org/abs/2512.02924
3. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
4. Daubechies, I.: Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1992). https://doi.org/10.1137/1.9781611970104
5. Fan, M., Geng, B., Li, K., Wang, X., Varshney, P.K.: Interpretable data fusion for distributed learning: A representative approach via gradient matching. In: 2024 27th International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2024)
6. Fan, M., Li, K., Zhang, T., Tian, Q., Geng, B.: Pfeddst: Personalized federated learning with decentralized selection training. In: 2025 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2025)
7. Fan, M., Zhang, T., Geng, B.: A unified framework for the convergence and weight pruning in federated learning. In: Proceedings of the 7th ACM International Conference on Multimedia in Asia. pp. 1–5 (2025)
8. Fan, M., Zhang, T., Ma, X., Guo, J., Zhan, Z., Zhou, S., Qin, M., Ding, C., Geng, B., Fardad, M., et al.: A unified dnn weight compression framework using reweighted optimization methods. Intelligent Systems with Applications p. 200556 (2025)

9. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. Journal of the Optical Society of America A **4**(12), 2379–2394 (1987). https://doi.org/10.1364/JOSAA.4.002379
10. Hernandez Petzsche, M.R., Kiesel, S., Huber, F., Labuschagne, C.A., Zamboni, N.S., Noth, J., Zietz, K., Buerger, M.G., Kaesmacher, J., Wiest, R., Reyes, M.: Public dataset of ischemic stroke lesion segmentation from multi-modal mri. Scientific Data **9**, 762 (2022). https://doi.org/10.1038/s41597-022-01875-5
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015), https://arxiv.org/abs/1503.02531
12. 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)
13. 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)
14. Lan, Q., Hsu, Y.C., Khan, N.S., Jiang, X.: Reco-kd: Region-and context-aware knowledge distillation for efficient 3d medical image segmentation. arXiv preprint arXiv:2601.08301 (2026)
15. 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)
16. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE transactions on pattern analysis and machine intelligence **11**(7), 674–693 (1989)
17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
18. Mezrich, R.: A perspective on k-space. Radiology **195**(2), 297–315 (1995)
19. Moratal, D., Vallés-Luch, A., Martí-Bonmatí, L., Brummer, M.E.: k-space tutorial: an mri educational tool for a better understanding of k-space. Biomedical imaging and intervention journal **4**(1), e15 (2008)
20. Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: International conference on machine learning. pp. 5301–5310. PMLR (2019)
21. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: International Conference on Learning Representations (ICLR) (2015)
22. 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)
23. Ruderman, D., Bialek, W.: Statistics of natural images: Scaling in the woods. Advances in neural information processing systems **6** (1993)
24. 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. pp. 5311–5320 (2021)
25. Unser, M., Aldroubi, A.: A review of wavelets in biomedical applications. Proceedings of the IEEE **84**(4), 626–638 (1996). https://doi.org/10.1109/5.488704
26. de Verdier, M.C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., et al.: The 2024 brain tumor

segmentation (brats) challenge: Glioma segmentation on post-treatment mri. arXiv preprint arXiv:2405.18368 (2024)
27. Wang, Y., Blackie, L., Miguel-Aliaga, I., Bai, W.: Memory-efficient segmentation of high-resolution volumetric microct images. In: International Conference on Medical Imaging with Deep Learning. pp. 1322–1335. PMLR (2022)
28. Wang, Y., Zhou, W., Jiang, T., Bai, X., Xu, Y.: Intra-class feature variation distillation for semantic segmentation. In: European Conference on Computer Vision. pp. 346–362. Springer (2020)
29. Xu, Y., Raj, A., Victor, J.D.: Systematic differences between perceptually relevant image statistics of brain mri and natural images. Frontiers in neuroinformatics **13**, 46 (2019)
30. Xu, Z.Q.J., Zhang, Y., Luo, T.: Overview frequency principle/spectral bias in deep learning. Communications on Applied Mathematics and Computation **7**(3), 827–864 (2025)
31. Xu, Z.Q.J., Zhang, Y., Luo, T., Xiao, Y., Ma, Z.: Frequency principle: Fourier analysis sheds light on deep neural networks. arXiv preprint arXiv:1901.06523 (2019)
32. Zhang, L., Chen, X., Tu, X., Wan, P., Xu, N., Ma, K.: Wavelet knowledge distillation: Towards efficient image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12464–12474 (2022)
33. Zhang, R.: Making convolutional networks shift-invariant again. In: International conference on machine learning. pp. 7324–7334. PMLR (2019)
34. Zhang, Y., Huang, T., Liu, J., Jiang, T., Cheng, K., Zhang, S.: Freekd: Knowledge distillation via semantic frequency prompt. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15931–15940 (2024)