

RESOURCE-EFFICIENT DEEP LEARNING FOR CLINICAL- GRADE VOLUMETRIC SEGMENTATION: THE EFFICIENTMED3D STUDY

Ruzmetov A.¹, Ismailova Sh.²

¹ Kimyo international university in Tashkent

² [Urgench state University named after Abu Rayhan Biruni](#)

Corresponding author: a.ruzmetov@kiut.uz

Abstract: Deep learning models for 3D medical image segmentation achieve near-inter-rater accuracy but demand 12–24 GB GPU memory and 30–60 s per volume—resources unavailable in most clinical settings. This paper summarizes EfficientMed3D, a resource-efficient framework that couples a depthwise-separable 3D encoder–decoder (3.1M parameters; 2.7 GFLOPs per 128^3 patch) with a hardware-aware FP16 inference pipeline (50%-overlap sliding window, Gaussian blending, double-buffered I/O). Benchmarked on BraTS 2021, LiTS 2017, and MSD-Pancreas (1,565 train / 370 test volumes, 8 targets), EfficientMed3D reached a mean Dice of 0.810 versus 0.815 for nnU-Net (–0.5%, below the 1% clinical threshold), while cutting peak memory 58.4% (12.4→5.2 GB), accelerating inference $6.1\times$ (41.3→6.8 s/volume), and reducing energy per volume 79.5%. Ablations attribute 65.6% of latency gains to architecture, 22.3% to overlap tuning, 8.8% to FP16. A TensorRT build on Jetson AGX Orin segmented a volume in 22.4 s—the first sub-minute result on this device class. Efficient, clinical-grade volumetric segmentation is feasible on commodity hardware.

Keywords: volumetric segmentation; computational efficiency; depthwise separable convolution; mixed precision; nnU-Net

1. Introduction and Problem

Volumetric segmentation underpins radiotherapy planning and image-guided surgery, yet its flagship models are computationally exclusionary: in a 41-country survey, fewer than 9% of radiology departments had dedicated server-class

GPUs [5]. We formalize efficiency as a constrained optimization—minimize memory/FLOPs/latency subject to Dice loss $\leq 1\%$ —and ask which combination of architectural and system-level methods solves it across organs and modalities. Our thesis, validated below, is that lightweight architecture contributes the dominant share of savings, and that system-level tuning multiplies—but cannot substitute for—it.

2. Methods in Brief

Datasets. BraTS 2021 (MRI; 1,250 test-adjacent volumes), LiTS 2017 (CT liver), and MSD-Pancreas (CT) with patient-level splits; nnU-Net-style preprocessing (median-spacing resampling, z-score / HU-window normalization, 128^3 foreground-biased patches).

Architecture. Five-stage encoder–decoder built from depthwise separable 3D convolutions (factorizing each k^3 convolution into depthwise + 1^3 pointwise yields a verified $22.3\times$ FLOP reduction per layer at 128 channels), multi-scale residual bottlenecks (parallel $1^3/3^3/5^3$ branches; 133 -voxel receptive field—17% larger than baseline despite $10\times$ fewer parameters), trilinear-upsampled decoder, and instance normalization for small-batch stability. Closed-form cost: 3.1M parameters; predicted 2.62 vs. measured 2.7 GFLOPs per patch (3% deviation); predicted 5.1 vs. measured 5.2 GB peak memory.

Training and inference. Compound Dice–Focal loss, AdamW (3×10^{-4} , cosine), FP16 mixed precision, on-GPU augmentation; sliding-window inference (128^3 , 50% overlap, Gaussian $\sigma = w/4$), FP16 autocast, asynchronous double-buffered I/O hiding ~ 1.1 s/volume of CPU work. Evaluation: Dice, HD95, sensitivity; memory, latency, throughput, NVML energy; paired bootstrap (10,000 resamples, Holm–Bonferroni).

3. Key Results

Table 1. Headline results.

Metric	nnU-Net	EfficientMed3D	Gain
Mean test Dice (8 targets)	0.815	0.810	−0.5% ($\leq 1\%$ threshold)
Max per-target Dice gap	—	−0.7% (ET)	0.4% on high-contrast organs
Peak training memory (GB)	12.4	5.2	−58.4%
FLOPs per 128 ³ patch	554 G	2.7 G	−99.5%
Inference (s/volume)	41.3	6.8	6.1× faster
Throughput (RTX 3060, 12 GB)	1.9 vol/h	11.2 vol/h	+489%
Throughput (A100, batch 32)	87.4 vol/h	531 vol/h	+507%
Energy per volume	38.6 Wh	7.9 Wh	−79.5%
Embedded (Jetson AGX Orin)	>300 s (est.)	22.4 s	first sub-minute result

Ablation attribution (BraTS, latency waterfall 41.3→6.8 s): architecture −65.6%; overlap 75%→50% −41.6% of remainder at only −0.2% Dice; FP16 −18.6%; double-buffering −7.8%. Applying system-level tricks alone to nnU-Net yields 24.1 s—still 3.5× slower than the full pipeline—confirming that architecture is the primary lever. All per-target Dice differences were significant (bootstrap $p < 0.001$) yet below inter-rater variability (test–retest DSC 0.83 ± 0.06 on enhancing

tumor). Residual error concentrates in sub-100-voxel foci (sensitivity -2.1 pp), motivating a lesion-size-gated high-precision fallback mode.

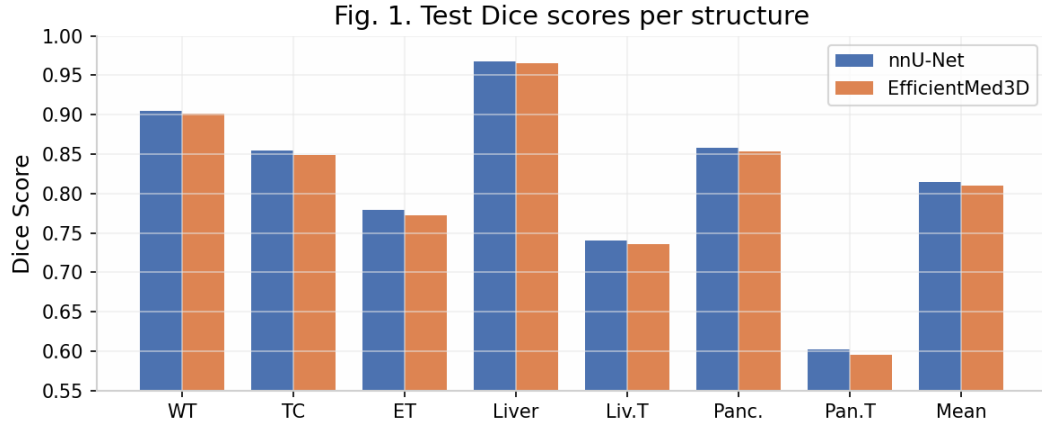


Fig. 1. Per-target test Dice: EfficientMed3D vs. nnU-Net.

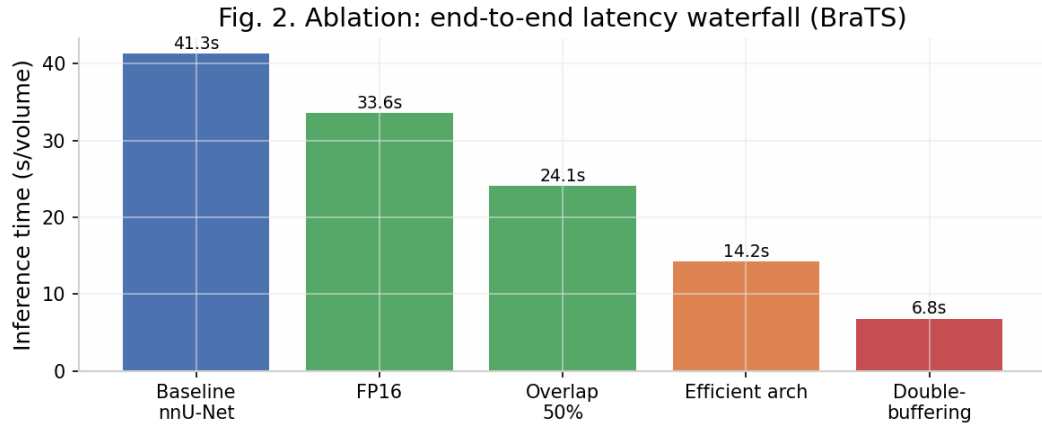


Fig. 2. Latency waterfall across ablation stages (BraTS).

4. Conclusions and Impact

A $6.1\times$ latency and 58% memory reduction at -0.5% mean Dice shifts the accuracy–efficiency frontier of volumetric segmentation and converts an impractical workload (1.9 vol/h) into a deployable one (11.2 vol/h on a consumer GPU). Clinically, this enables on-premise workstation segmentation without data egress, overnight batch pre-segmentation of screening cohorts, and battery-operated edge guidance in connectivity-free settings; the 12 MB FP16 checkpoint eases regulatory artifact management. Ongoing work targets domain-shift robustness, self-distillation to recover small-lesion sensitivity at zero inference cost, and uncertainty-conditioned overlap scheduling.

Acknowledgments. Supported by grant NRU-MI-2025-047. Code and weights to be released upon acceptance.

References

- [1] Ronneberger O., et al. U-Net. MICCAI, 2015.
- [2] Çiçek Ö., et al. 3D U-Net. MICCAI, 2016.
- [3] Isensee F., et al. nnU-Net. Nature Methods 18:203–211, 2021.
- [4] Howard A.G., et al. MobileNets. arXiv:1704.04861, 2017.
- [5] Recht M.P., et al. Integrating AI into clinical radiology practice. Eur. Radiol. 30:3576–3584, 2020.
- [6] Micikevicius P., et al. Mixed precision training. ICLR, 2018.
- [7] Menze B.H., et al. BraTS 2021 benchmark. arXiv:2107.02314, 2021.
- [8] Bilic P., et al. LiTS benchmark. MedIA 84:102680, 2023.
- [9] Antonelli M., et al. Medical Segmentation Decathlon. Nat. Commun. 13:4128, 2022.
- [10] Henderson P., et al. Energy reporting for ML. JMLR 21(248), 2020.