OmniSegNet: Towards Scalable, Efficient & Universal Medical Image Segmentation

Soma Dasgupta, Swarnava Dey, Arijit Mukherjee and Arpan Pal

Problem statement & challenge

Problem:
Despite advances in segmentation, current architectures fail to deliver a unified, efficient solution adaptable to multiple medical imaging modalities and constrained hardware environments, hindering real-world clinical deployment.

Use cases:

- Enables lightweight, edge-deployable medical image segmentation across modalities like skin, organ, retinal, and surgical imaging for real-time clinical inference.

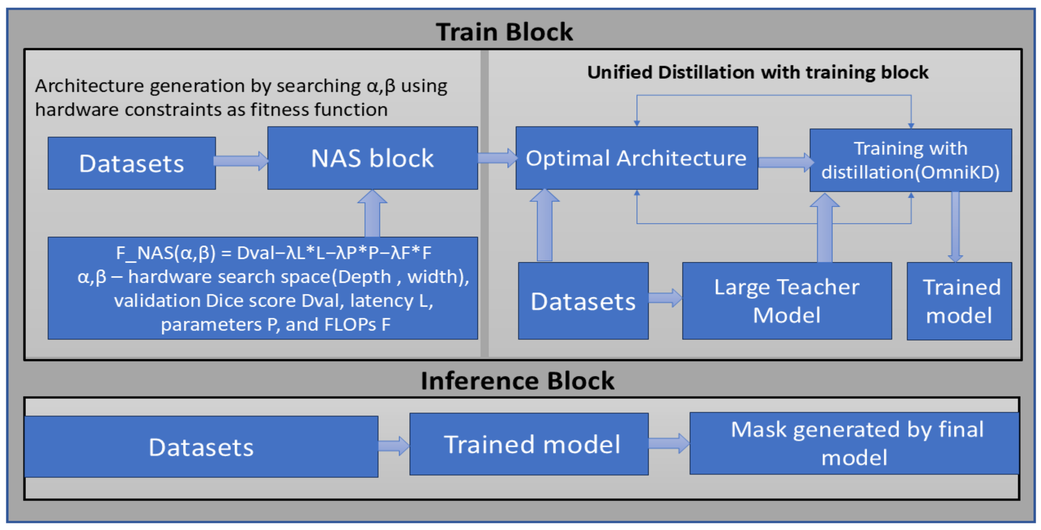
Challenges:

- **Cross-modality gap:** Models struggle to generalize across diverse medical imaging modalities.
- **Edge inefficiency:** Large models are unsuitable for real-time, privacy-preserving edge deployment.
- **Manual tuning burden:** Achieving accuracy, efficiency balance requires costly task-specific optimization in automated manner.

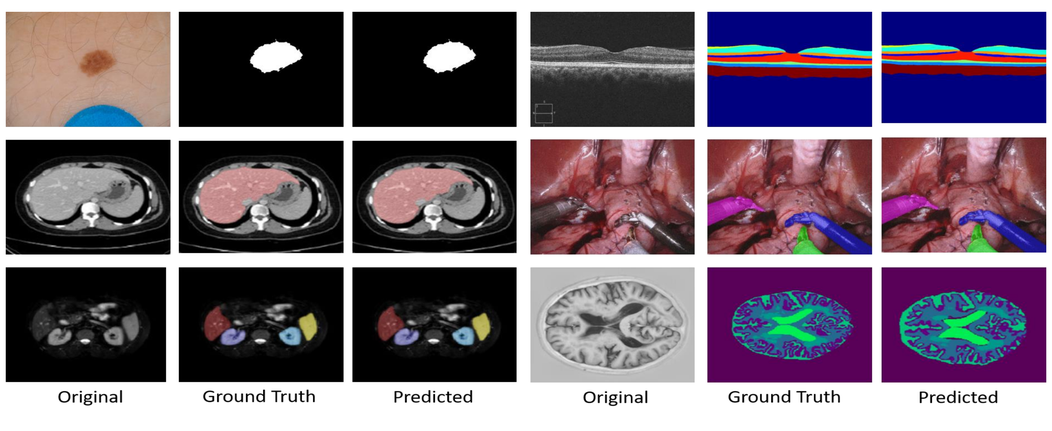
How to build one model that works efficiently across all modalities and edge devices?

- **Unified Architecture:** Scalable design generalizes across medical modalities.
- **Edge Optimization:** Latency-aware pareto-optimal NAS that ensures efficient architectures for diverse hardware budget.
- **Unified Distillation:** Multi-term distillation loss, OmniKD that enables dataset-agnostic knowledge transfer for edge architecture optimization.

Proposed Workflow



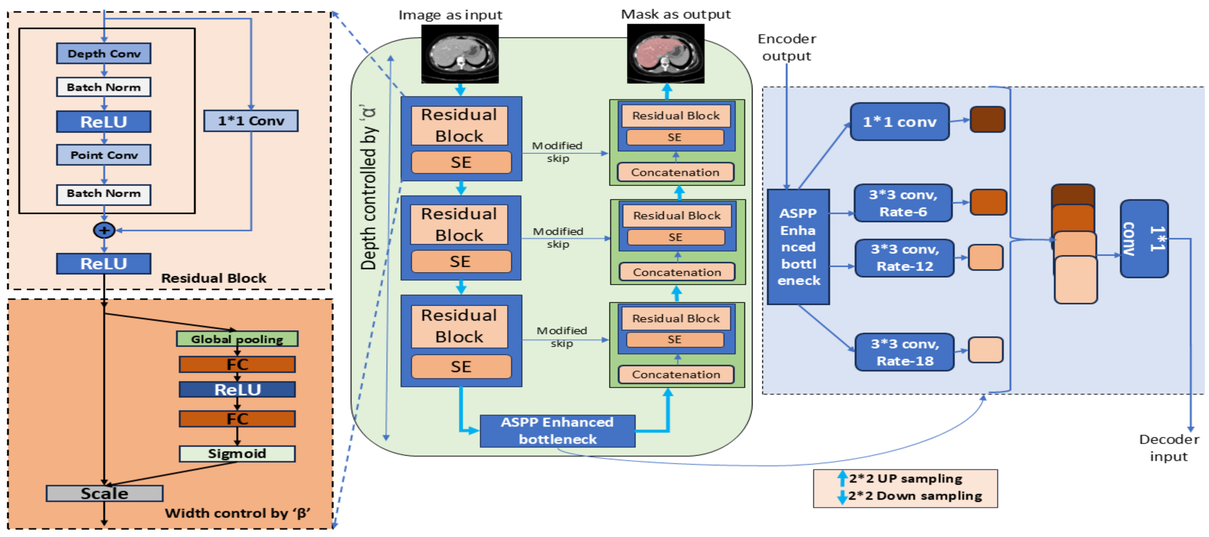
Mask generation by Trained OmniSegNet



Observations on generated mask

- The trained model produces **high-quality segmentation masks** with sharp, anatomically consistent boundaries across modalities.
- **SE-enhanced encoder** focuses on salient structures while suppressing noise, and **ASPP bottleneck** captures diverse object scales effectively.
- **Modified skip fusion and SE-decoder** refine spatial details, ensuring smooth, continuous contours without fragmentation.
- Through **OmniKD distillation**, the model retains semantic depth and contextual awareness comparable to large teachers like SAM.
- Resulting masks show **strong generalization and clinical reliability**, even under constrained edge-device settings.

Overall architecture and workflow



Architecture and Methodology

1. Latency-Aware NAS (Optimization Stage)

- Evolutionary compound-scaling search (α, β) discovers Pareto-optimal models balancing Dice accuracy, parameters, and edge latency.
- Automatically adapts model size for devices from Raspberry Pi to Jetson Orin without manual tuning.

2. Scalable Architecture (Design Stage)

- SE-enhanced residual encoder + ASPP bottleneck + modified skip decoder for efficient local-global context fusion.
- Compound depth–width scaling provides Compact, Balanced, and Accurate variants under a single unified backbone.

3. Unified Distillation (Training Stage)

- OmniKD fuses five cues — logits, features, attention, relational, contextual — from fine-tuned SAM teacher models.
- Enables modality-agnostic supervision and eliminates dataset-specific loss engineering.

Results and Takeaways

Dataset	OmniSegNet-C DSC	OmniSegNet-C IoU	OmniSegNet-C Lat. (ms)	OmniSegNet-C Params (M)	CMUNeXt [3] DSC	CMUNeXt [3] IoU	CMUNeXt [3] Lat.	CMUNeXt [3] Params	YOLOv11n-Seg [1] DSC	YOLOv11n-Seg [1] IoU	YOLOv11n-Seg [1] Lat.	YOLOv11n-Seg [1] Params
ISIC 2017	**0.87±0.01**	**0.80±0.01**	**24**	**1.5**	0.80±0.02	0.73±0.02	35	0.9	0.81±0.01	0.74±0.01	38	3.0
CHAOS CT	**0.89±0.01**	**0.81±0.01**	**25**	**1.6**	0.81±0.02	0.73±0.02	34	1.1	0.83±0.01	0.75±0.01	39	3.1
CHAOS MRI	**0.85±0.01**	**0.79±0.01**	**30**	**1.6**	0.78±0.02	0.70±0.02	33	0.8	0.80±0.02	0.72±0.02	40	2.9
Retinal OCT	**0.89±0.01**	**0.82±0.01**	**16**	**1.0**	0.84±0.02	0.76±0.02	24	0.7	0.86±0.01	0.78±0.01	28	2.9
Intra-op	**0.86±0.01**	**0.81±0.01**	**22**	**1.3**	0.80±0.02	0.71±0.02	30	0.9	0.82±0.01	0.74±0.01	33	3.0
MRBrainS18	**0.86±0.01**	**0.81±0.01**	**27**	**1.6**	0.77±0.02	0.68±0.02	28	0.8	0.79±0.01	0.70±0.01	30	3.1
Dataset	**OmniSegNet-B DSC**	**OmniSegNet-B IoU**	**OmniSegNet-B Lat. (ms)**	**OmniSegNet-B Params (M)**	**CMUNeXt [3] DSC**	**CMUNeXt [3] IoU**	**CMUNeXt [3] Lat.**	**CMUNeXt [3] Params**	**YOLOv8n-Seg [2] DSC**	**YOLOv8n-Seg [2] IoU**	**YOLOv8n-Seg [2] Lat.**	**YOLOv8n-Seg [2] Params**
ISIC 2017	**0.89±0.01**	**0.81±0.01**	**32**	**3.1**	0.82±0.01	0.75±0.01	40	3.2	0.83±0.01	0.76±0.01	42	3.5
CHAOS CT	**0.89±0.01**	**0.82±0.01**	**33**	**3.2**	0.83±0.01	0.75±0.01	38	3.3	0.85±0.01	0.77±0.01	41	3.5
CHAOS MRI	**0.88±0.01**	**0.84±0.01**	**40**	**3.3**	0.80±0.01	0.72±0.01	39	3.2	0.82±0.01	0.74±0.01	43	3.4
Retinal OCT	**0.90±0.01**	**0.86±0.01**	**22**	**1.8**	0.85±0.01	0.77±0.01	28	3.1	0.87±0.01	0.78±0.01	30	3.5
Intra-op	**0.89±0.01**	**0.85±0.01**	**28**	**2.5**	0.82±0.01	0.74±0.01	33	3.2	0.84±0.01	0.76±0.01	37	3.4
MRBrainS18	**0.89±0.01**	**0.85±0.01**	**31**	**3.0**	0.80±0.01	0.71±0.01	35	3.3	0.82±0.01	0.73±0.01	36	3.4
Dataset	**OmniSegNet-A DSC**	**OmniSegNet-A IoU**	**OmniSegNet-A Lat. (ms)**	**OmniSegNet-A Params (M)**	**CMUNeXt-L [3] DSC**	**CMUNeXt-L [3] IoU**	**CMUNeXt-L [3] Lat.**	**CMUNeXt-L [3] Params**	**FastSAM [4] DSC**	**FastSAM [4] IoU**	**FastSAM [4] Lat.**	**FastSAM [4] Params**
ISIC 2017	**0.92±0.01**	**0.89±0.01**	**38**	**6.0**	0.83±0.01	0.76±0.01	60	8.1	0.84±0.01	0.77±0.01	65	10.2
CHAOS CT	**0.93±0.01**	**0.91±0.01**	**40**	**6.2**	0.85±0.01	0.77±0.01	58	8.3	0.87±0.01	0.79±0.01	62	10.2
CHAOS MRI	**0.96±0.01**	**0.92±0.01**	**42**	**6.2**	0.83±0.01	0.74±0.01	57	8.2	0.84±0.01	0.76±0.01	61	10.1
Retinal OCT	**0.96±0.01**	**0.91±0.01**	**26**	**3.0**	0.87±0.01	0.79±0.01	35	8.1	0.88±0.01	0.80±0.01	40	10.2
Intra-op	**0.94±0.01**	**0.91±0.01**	**30**	**4.2**	0.86±0.01	0.78±0.01	50	8.2	0.87±0.01	0.79±0.01	55	10.2
MRBrainS18	**0.96±0.01**	**0.87±0.01**	**35**	**4.7**	0.83±0.01	0.75±0.01	52	8.3	0.85±0.01	0.76±0.01	58	10.1

Table 1: Comparison of OmniSegNet variants with recent models (CMUNeXt [3], YOLOv11n-Seg [1], YOLOv8n-Seg [2], CMUNeXt-L [3], and FastSAM [4]) across six medical segmentation datasets. The three blocks correspond to Compact, Balanced, and Accurate variants, respectively. Reported values are mean±std over 3-fold CV. Bold indicates the best mean. Significance was assessed using paired two-sided t-tests over CV folds with Benjamini–Hochberg FDR correction (q=0.05)

- **Higher Accuracy:** OmniSegNet consistently achieves 6–12% higher Dice and IoU scores than CMUNeXt, YOLOv8/11n-Seg, and FastSAM across all datasets.
- **Lower Latency:** It delivers real-time segmentation with **24–40 ms inference** while keeping parameters between **1–6 M**.
- **Better Scalability:** Compact, Balanced, and Accurate variants maintain strong **accuracy–efficiency trade-offs**, proving adaptability across diverse modalities and hardware.

Future Works

- Extend OmniSegNet to **3D and temporal medical data** for volumetric and video-based segmentation.
- Integrate **self-supervised and continual learning** for adaptive edge inference across unseen modalities and clinical domains.

References
[1] R. Khanam and M. Hussain, *"YOLOv11: An overview of the key architectural enhancements,"* arXiv preprint **arXiv:2410.17725**, 2024. [Online]. Available: https://arxiv.org/abs/2410.17725
[2] M. Sohan, T. S. Ram, and C. V. R. Reddy, *"A review on YOLOv8 and its advancements,"* in *Proc. Int. Conf. on Data Intelligence and Cognitive Informatics*, Springer, 2024, pp. 529–545.
[3] F. Tang, J. Ding, Q. Quan, L. Wang, C. Ning, and S. K. Zhou, *"CMUNeXt: An efficient medical image segmentation network based on large kernel and skip fusion,"* in *Proc. IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 2024, pp. 1–5.
[4] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, *"Fast Segment Anything,"* arXiv preprint **arXiv:2306.12156**, 2023.