OmniSegNet: Towards Scalable, Efficient & Universal Medical Image Segmentation

Soma Dasgupta
dasgupta.soma@tcs.com
Swarnava Dey
swarnava.dey@tcs.com
Arijit Mukherjee
mukherjee.arijit@tcs.com
Arpan Pal
arpan.pal@tcs.com

TCS Research
Tata Consultancy Services
Kolkata, India

Abstract

Medical image segmentation spans diverse modalities, including MRI (Magnetic Resonance Imaging), CT (Computed Tomography), OCT (Optical Coherence Tomography), and USG (Ultrasound Sonography), each with unique spatial and contextual characteristics. This heterogeneity demands architectures that balance global context with fine-grained anatomical details, a challenge for standard models like U-Net and Transformer-based approaches, which either lack generalization or struggle with subtle features. Privacy constraints further restrict the use of large models like the Segment Anything Model (SAM) for cloud-based inference, complicating edge deployment and increasing development complexity.

To address these challenges, we propose **OmniSegNet**, a unified, scalable encoder-decoder architecture that integrates SE-enhanced residual blocks for efficient local-global context capture and Atrous Spatial Pyramid Pooling (ASPP) for multi-scale feature aggregation. Compound scaling of depth (α) and width (β) supports flexible model variants ranging from 1.5M to 6M parameters, identified through latency-aware neural architecture search for real-time deployment on devices like Raspberry Pi and Jetson Nano.

To generalize without task-specific tuning, we introduce **OmniKD**, a unified distillation framework that transfers knowledge from fine-tuned SAM models via logits, intermediate features, attention maps, relational structures, and contextual similarities, eliminating the need for handcrafted loss functions.

OmniSegNet achieves up to 10× parameter reduction while improving Dice scores by 10–12% across diverse medical imaging benchmarks, including ISIC, CHAOS, and MRBrainS18, offering a scalable, efficient, and privacy-preserving solution for real-world medical segmentation.

1 Introduction

Medical image segmentation is fundamental for numerous clinical tasks, including diagnosis, intervention planning, and treatment evaluation. It spans a wide range of imaging modalities, such as CT (Computed Tomography), MRI (Magnetic Resonance Imaging), ultrasound, and histopathology, each characterized by unique spatial resolutions and tissue contrasts. These modality-specific variations often require tailored architectures and specialized training procedures, as a model optimized for liver segmentation on CT may not generalize well to detecting brain lesions on MRI. Moreover, the scale of target structures can vary dramatically—from tiny melanoma spots in dermoscopy to large anatomical regions in abdominal CT scans—posing additional challenges for universal model design. These challenges typically necessitate handcrafted and specialized architectures for each dataset-task combination, significantly increasing development complexity and deployment costs.

The primary architectures for medical segmentation are based on convolutional neural networks (CNNs), particularly U-Net [28] and its numerous variants [1, 5, 10, 23]. These models have demonstrated remarkable performance on individual tasks but suffer from several limitations when generalized across diverse medical datasets. For instance, extensions like UNet++ [45] and TransUNet [7] introduce nested skip connections and Transformer-CNN hybrids, respectively, to improve long-range context modeling. However, these approaches still rely on manually curated datasets and handcrafted pipelines, making cross-domain adaptation challenging. Transformers, despite their ability to capture long-range dependencies, often struggle to preserve fine-grained local details, which are critical for accurate medical segmentation. This limitation is particularly problematic in medical images, where the foreground-to-background ratio is typically low, and subtle anatomical boundaries must be accurately delineated.

Beyond these architectural challenges, recent foundation models like SAM [20] and MedSAM [24] have emerged with the aim of providing generalized, cross-domain segmentation capabilities. However, these models tend to underperform on low-contrast medical images without extensive fine-tuning, limiting their out-of-the-box applicability. Moreover, their high computational requirements make them unsuitable for real-time on-premises medical image analytics, where latency and power consumption are critical. This introduces a second significant challenge: the need for privacy-preserving, on-device inference. Given the sensitivity of medical data, cloud-based processing is often prohibited, necessitating efficient, locally deployable models that can operate within the constrained environments of standard personal computers / workstations in the hospital, not necessarily with a high-end GPU. Building such models requires not only deep learning expertise but also medical domain-specific knowledge and multi-objective optimization skills, resources that are rarely available at scale.

To address these challenges, we propose **OmniSegNet**, a unified, scalable encoder-decoder architecture that generalizes across diverse modalities, organs, and deployment platforms. OmniSegNet integrates squeeze-and-excitation (SE) enhanced residual blocks [14] for effective local-global context capture and utilizes Atrous Spatial Pyramid Pooling (ASPP) [8] for multi-scale feature aggregation, essential for accurate segmentation across a wide range of targets. The architecture supports compound scaling through depth (α) and width (β) parameters [32], allowing for flexible model variants tailored to resource-constrained devices.

To automate model design and reduce manual tuning, we employ a latency-aware evolutionary neural architecture search (NAS), inspired by MnasNet [33] and ProxylessNAS [3]. This approach identifies Pareto-optimal trade-offs between latency and accuracy, significantly reducing the time and expertise required for model design while ensuring efficient, real-time performance on edge hardware.

To further enhance generalization without relying on task-specific loss functions, we introduce **OmniKD**, a unified knowledge distillation framework. OmniKD transfers knowledge from a fine-tuned SAM teacher through multiple complementary modes, including soft targets (logits), intermediate features [29], attention maps [41], relational structures [27], and contextual similarities [22]. This multi-term distillation strategy allows small models to inherit rich supervision from large, domain-specific teachers without requiring dataset-specific tuning. We distinguish between *cross-modality* generalization, which spans distinct imaging physics such as CT, MRI, OCT, and dermoscopy, and *cross-domain* generalization, which ensures robustness within a modality across scanners or acquisition protocols. OmniSegNet tackles both by combining SE-enhanced residual blocks with ASPP and compound scaling to adapt across modalities, while OmniKD transfers modality-invariant and domain-robust cues from SAM, enabling compact students to generalize without dataset-specific tuning.

Key Contributions: We introduce **OmniSegNet**, a unified encoder-decoder architecture with SE-enhanced residual blocks and ASPP for efficient multi-scale context capture, supporting compound scaling (α–β) for flexible model variants (1.5M–10M parameters) on edge devices. We employ latency-aware NAS for Pareto-optimal design under real-time constraints and propose **OmniKD**, a unified distillation framework integrating logits, features, attention, relational cues, and contextual similarity, enabling compact models to leverage rich supervision from large foundation models without dataset-specific tuning.

2 Related Work

Existing segmentation approaches in the medical domain range from classical U-Net [28] variants to large-scale promptable foundation models like SAM [20] and MedSAM [24]. While each method achieves task-specific gains, most lack scalability across modalities, require high manual effort, or are unsuitable for low-resource edge deployment.

U-Net [28] remains the cornerstone of biomedical segmentation due to its encoder–decoder design and skip connections. Enhancements like UNet++ [45] and TransUNet [7] improve upon it by narrowing the semantic gap and modeling long-range dependencies using Transformer based encoders. SegFormer [38] further extends this paradigm with hierarchical attention modules for efficient multi-scale processing. However, these architectures require modality-specific tuning and incur significant computational overhead, which is not suitable for resource-constrained edge devices. For efficiency, ESPNet [25] employs an Efficient Spatial Pyramid (ESP) to lower computation while preserving accuracy though it lacks in generalization, across different modalities across medical images. To enhance generalization our architecture adopts a U shaped encoder decoder backbone enhanced with residual blocks, Squeeze-and-Excitation (SE) [14], and Atrous Spatial Pyramid Pooling (ASPP) [8], enabling scalable and efficient local-global feature fusion. Foundation models like SAM [20] and MedSAM [24] offer promptable segmentation using large-scale training but demand heavy computation and are limited to prompt-driven settings. Lightweight adaptations like MobileSAM [39, 43] compress the encoder for mobile platforms but still rely on prompt-based inference. Our framework diverges by enabling prompt-free, unified training across modalities on resource-constrained devices.

Neural Architecture Search (NAS) methods such as EfficientNet [32] introduced compound scaling to balance depth, width, and resolution, while Once-for-All (OFA) [4] enabled a supernet to support diverse sub-networks. MnasNet [33] and ProxylessNAS [3] integrated latency into the search objective but often relied on proxy metrics or specific hardware assumptions. In contrast, our evolutionary NAS framework(Refer Sub Section- 3.2) performs direct search on device latency and parameters using multi-objective optimization over the compound scaling space (α, β). This ensures deployment-aligned, Pareto-optimal architectures tailored to both accuracy and runtime efficiency under practical edge constraints.

Knowledge distillation (KD) has advanced from soft logits [13] to richer forms including features [29], attention [41], relational [26], and contextual cues [35]. However, existing methods often require domain-specific tuning or lack scalability. Our proposed **OmniKD** unifies all five strategies into a single framework, enabling compact models to learn from large teachers across diverse modalities (CT, MRI, dermoscopy). As Table 2 shows, this delivers prompt-free, efficient, and generalizable segmentation with consistent performance across multi-domain medical imaging tasks.

3 Methodology

3.1 OmniSegNet: A Scalable U-Net Backbone

OmniSegNet is a fully convolutional encoder-decoder architecture tailored for scalable, multi-domain medical image segmentation. Its structural complexity is modulated by compound scaling parameters: α (depth) and β (width). These parameters jointly control the number of encoding/decoding stages and feature channel dimensionality, enabling flexible trade-offs between accuracy, model size, and latency across deployment scenarios.(Refer Table 1.)

As depicted in Figure 1, the encoder comprises sequential SE-enhanced residual blocks. Each block integrates a depthwise convolution, batch normalization, ReLU activation, followed by a pointwise convolution. A skip connection adds the input to the output of the block to maintain gradient flow. The output is then passed to a Squeeze-and-Excitation (SE) module which performs global average pooling, two fully connected layers, and a sigmoid activation. This generates channel-wise attention coefficients that rescale the input, effectively suppressing irrelevant features and enhancing salient activations. Downsampling between encoder stages is performed using 2×2 max pooling layers with stride 2, which simultaneously reduce spatial resolution and enlarge the receptive field.

At the bottleneck, OmniSegNet integrates an Atrous Spatial Pyramid Pooling (ASPP) module to encode multi-scale context. It consists of parallel atrous convolutions with dilation rates $\{1, 6, 12, 18\}$ and a global average pooling branch. These outputs are concatenated and fused through a 1×1 convolution to produce a rich semantic representation.

The decoder takes up-sampled features from the previous block and concatenates them with skip features from SE-enhanced residual encoder blocks. This fused tensor, termed as a *modified skip*, is passed through a concatenation block that applies group convolutions for separate feature extraction, followed by two inverted bottleneck point wise convolutions. This enables adaptive and expressive fusion before feeding into the SE-enhanced residual decoder block, ensuring spatial detail restoration and efficient feature refinement across decoding stages. The concatenated feature map is passed through an SE-enhanced residual block that not only refines semantic boundaries, but also regulates the channel dimensionality. Specifically, the SE module selectively re-weights and prunes less informative channels before residual transformation, maintaining compact feature representations while preserving relevant context. This promotes efficient decoding by reducing redundant computation and guiding the model to focus on semantically meaningful activations. Detailed mathematical formulations and derivations of the proposed architecture explained in the supplementary material for comprehensive understanding.

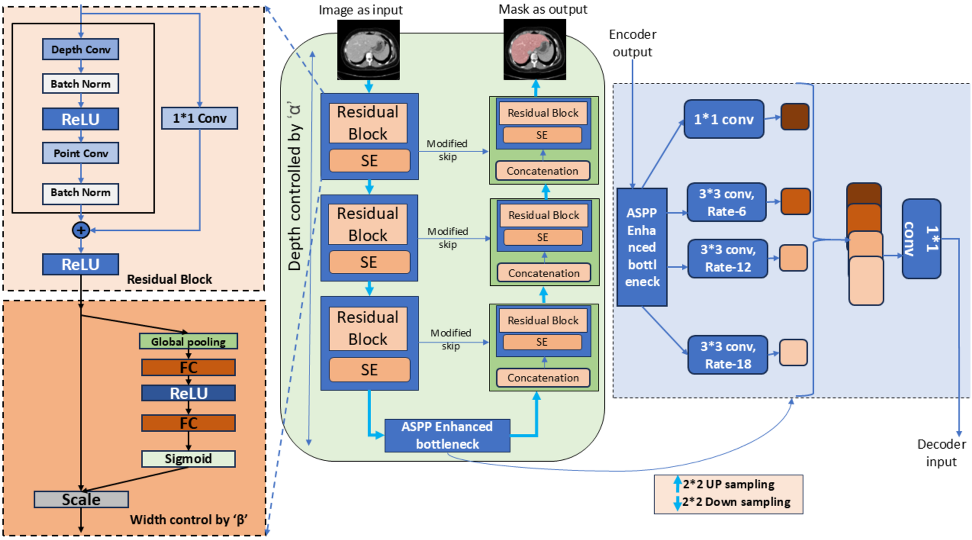

Figure 1: OmniSegNet: our proposed framework. SE: Squeeze and Excitation module.

3.2 Compound NAS for OmniSegNet

OmniSegNet introduces a novel Neural Architecture Search (NAS) strategy based on compound scaling parameters α and β, which act as global depth and width multipliers. Unlike traditional NAS [1, 16, 36] methods that tune low-level operations per layer, our approach searches a continuous, interpretable space where each candidate model $\mathcal{A}_{\alpha,\beta}$ is instantiated by scaling the backbone using (α,β) with $\alpha \in [0.6, 1.5]$ and $\beta \in [0.5, 2.0]$, constrained to $\mathbb{R}^+$. For each configuration, we compute validation Dice score D_{val}, latency L, parameters P, and FLOPs F, optimizing the multi-objective fitness $F_{\text{NAS}}(\alpha,\beta) = D_{\text{val}} - \lambda_L L - \lambda_P P - \lambda_F F$, where $\lambda_L, \lambda_P, \lambda_F$ balance accuracy and efficiency trade-offs effectively. This global scaling strategy enables rapid adaptation to diverse deployment constraints without costly, fine-grained per-layer tuning. We employ a modified NSGA-II algorithm, enhancing traditional Pareto-based search by introducing specialized crossover and mutation operators that respect continuous scaling, evolving architectures trained briefly on 20% of data to efficiently explore the search space. The NAS search was conducted on the ISIC2017 dataset (20%

Variant	α	β	#Enc Blks	#Dec Blks	Params(M)	FLOPs(B)	Deployment Target
Compact	0.62	0.54	6	6	1.5	4.8	**Ultra-low Edge** (Raspberry Pi 4, Jetson Nano)
Balanced	1.00	1.00	10	10	3.1	12.3	**Mid-tier Edge** (Jetson Xavier NX)
Accurate	1.38	1.60	14	14	6.0	22.0	**Efficient Edge** (Jetson Orin Nano, Google Coral)

Table 1: Final OmniSegNet variants discovered via compound NAS. Each variant reflects a globally scaled architecture tuned for specific hardware deployment scenarios.

subset), and the resulting Pareto-optimal variants (Table 1) were subsequently transferred and evaluated across all benchmarks.

3.3 OmniKD: Unified Multi-Term Knowledge Distillation

To bridge the performance gap between compact student models and large-capacity teacher models, we propose **OmniKD**, a unified knowledge distillation (KD) approach. Here, the teacher model is a fine-tuned version of SAM, which serves as the high-capacity model to guide the student. Instead of separately fine-tuning individual distillation losses for each dataset, OmniKD consolidates five complementary supervision signals into a single, dataset-agnostic distillation objective. This allows for better scalability across various datasets in medical imaging tasks.

Our unique search strategy employs Bayesian Optimization with Gaussian Processes to tune the loss weights λ for OmniKD, optimizing the trade-offs between performance and computational efficiency. This approach enables efficient distillation across multiple datasets, minimizing the need for dataset-specific fine-tuning.

The OmniKD loss is expressed as $\mathcal{L}_{\text{OmniKD}} = \sum_{i=1}^{5} \lambda_i \cdot \mathcal{L}_i$, where the components $\mathcal{L}_i$ include logits ($\mathcal{L}_{\text{logit}}$), features ($\mathcal{L}_{\text{feat}}$), attention maps ($\mathcal{L}_{\text{att}}$), relational geometry ($\mathcal{L}_{\text{rel}}$), and contextual embeddings ($\mathcal{L}_{\text{ctx}}$). The weights λ_i are optimized for each dataset using Bayesian Optimization to balance performance and computational efficiency.

For each of the six datasets, the optimal weight vector $\lambda = [\lambda_1, \ldots, \lambda_5]$ is selected to maximize the validation Dice score while minimizing latency, using the fitness function $F(\lambda) = \text{Dice}_{\text{val}}(\lambda) - \gamma \cdot \text{Latency}$. This approach eliminates the need for dataset-specific fine-tuning of each individual loss term, enabling efficient distillation across multiple domains. The Bayesian optimization adds a modest overhead: each dataset-specific search required approximately 1 GPU-hours, which is far lower than manual hyperparameter tuning or teacher training. Once identified, the optimal λ weights are fixed for deployment, incurring no additional runtime cost.

The optimized λ values for the six key datasets are presented in Table 3. Further details and configurations for other datasets are provided in the supplementary material.

Although Table 3 reports dataset-specific λ, we observed that a *single, universal* configuration (obtained by averaging the per-dataset optima) achieves performance within ∼2–4 pp Dice of the dataset-specific settings across all benchmarks in Section 4. This finding highlights that OmniKD can be deployed *without* per-dataset hyperparameter tuning, substantially reducing operational overhead and reinforcing its practicality for real-world medical imaging workflows.

Approach	**MDS**	**EI**	**PF**	**UKD**
U-Net/UNet++/TransUNet [7, 28, 45]	✗	✓	✗	✗
ESPNet [25]	✗	✓	✗	✗
SAM/MedSAM [20, 24]	✓	✗	✗	✗
MobileSAM/MobileSAMv2 [39, 43]	✗	✓	✗	✗
Domain-adaptive KD [6]	✓	✗	✓	✗
Ours (OmniSegNet + OmniKD)	✓	✓	✓	✓

Table 2: Comparison of MDS: Multi-domain Support, EI: Efficient Inference, PF: Prompt-free, UKD: Unified KD segmentation

Dataset(Ref.)	λ_{logit}	λ_{feat}	λ_{att}	λ_{rel}	λ_{ctx}
ISIC 2017 [9]	1.5	2.0	1.0	0.5	1.5
CHAOS CT [17]	2.0	2.0	1.0	0.5	1.0
CHAOS MRI [17]	1.5	1.5	1.0	0.5	1.0
Retinal OCT [18]	2.0	1.5	1.0	0.5	1.0
Intra-op [30]	1.5	1.5	1.0	0.5	1.5
MRBrainS18 [21]	1.8	2.5	1.0	0.5	1.0

Table 3: Optimized λ weights for OmniKD loss across 6 representative datasets (via Bayesian Optimization)

4 Experimental Results

4.1 Dataset Preparation and Augmentation

OmniSegNet is evaluated on six diverse medical imaging datasets, covering a broad spectrum of anatomical regions and imaging modalities, including dermatology, radiology, ophthalmology, and intra-operative imaging. The datasets, summarized in Table 4, are preprocessed to a unified resolution (256×256 or 512×512), followed by modality-specific normalization. Standard data augmentation techniques such as flips, deformations, affine transformations, and contrast adjustments are applied to all training samples, effectively doubling the size of the training set. SAM is fine-tuned separately in each dataset, ensuring a comprehensive evaluation in clinical domains. Detailed dataset-specific augmentation and preprocessing steps are provided in the supplementary material.

Dataset (Ref.)	**Task(Seg.)**	**Modality**	**Ch**	**Image Size**	**Train**	**Test**
ISIC 2017 [9]	Lesion	Dermoscopy	3	512×512	2000	600
CHAOS CT [17]	Organ	Abdominal CT	1	256×256	1120	280
CHAOS MRI [17]	Organ	Abdominal MRI	1	256×256	800	200
Retinal OCT [18]	Layer	Retinal OCT	1	512×512	3500	900
Intra-op [30]	Instruments	RGB Surgical	3	512×512	450	100
MRBrainS18 [21]	Brain Tissue	Brain MRI	3	240×240	144	192

Table 4: Dataset Summary: Task, Modality, Channels, Resolution, Train/Test Splits. Augmentations double the training set size.

4.2 Performance Comparison on Key Medical Datasets

We evaluate **OmniSegNet** in three configurations—**Compact (C)**, **Balanced (B)**, and **Accurate (A)**—across six diverse medical image segmentation datasets (see Table 4).

Each variant is compared against a common set of baseline models with similar parameter capacities, followed by state-of-the-art (SoTA) models that match in computational budget. Baseline models [19, 31, 32, 44] are carefully selected for their modular architecture, availability of lightweight variants, and structural consistency, enabling a fair and controlled comparison that highlights the strengths and adaptability of OmniSegNet.

All models were re-implemented and trained under identical conditions using consistent train/validation/test splits for each dataset. Architectures were adapted to accommodate dataset-specific input channels and segmentation mask formats. Training was conducted using PyTorch 2.2 with mixed precision on a system equipped with a 16GB NVIDIA RTX 3060 GPU, 16GB RAM, and an Intel i7 CPU. Optimization employed the AdamW optimizer with a learning rate of 3×10^{-4}, weight decay of 10^{-2}, and parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, alongside cosine annealing learning rate scheduling and early stopping. Each model was trained for 100 epochs using 3-fold cross-validation with a batch size of 8 on input images of size 256×256, augmented via random flips, rotations, scaling, and intensity jittering. Reported metrics correspond to the average performance across folds. Inference latency was measured on an edge-relevant CPU device (Dell Inspiron i5-10210U, 8GB RAM, PyTorch CPU backend), with latency ranging from 16 ms to 42 ms across the three OmniSegNet variants. Despite limited resources, all variants consistently outperformed size-matched baselines in both accuracy and latency (Table 5), demonstrating their suitability for edge deployment. *Statistical reporting.* We report mean±std across 3 folds and mark significant

Dataset	OmniSegNet-C				CMUNeXt [34]				YOLOv11n-Seg [19]			
	DSC	IoU	Lat. (ms)	Params (M)	DSC	IoU	Lat.	Params	DSC	IoU	Lat.	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**				**CMUNeXt [34]**				**YOLOv8n-Seg [31]**			
	DSC	IoU	Lat. (ms)	Params (M)	DSC	IoU	Lat.	Params	DSC	IoU	Lat.	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**				**CMUNeXt-L [34]**				**FastSAM [44]**			
	DSC	IoU	Lat. (ms)	Params (M)	DSC	IoU	Lat.	Params	DSC	IoU	Lat.	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 5: Comparison of OmniSegNet variants with recent models (CMUNeXt [34], YOLOv11n-Seg [19], YOLOv8n-Seg [31], CMUNeXt-L [34], and FastSAM [44]) 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) [2].

gains ($p<0.05$, paired two-sided t-test with BH correction [2]). See Supplementary for seeds and the exact test protocol. **Comparison with unconstrained baselines.** We further benchmark against large-scale architectures such as SegFormer-B3 [38], Swin-UNETR [12], and nnU-Net [15] (see Table 6). Despite their substantially larger capacity (>40M parameters), OmniSegNet-A (6M) achieves Dice within 1–2pp of these models, while offering up to 8–10× faster inference. This highlights that our compound-scaled backbone with OmniKD closes most of the accuracy gap at a fraction of the computational cost.

4.3 Distillation Effectiveness: OmniKD vs. Large Baselines

To narrow the gap between compact deployable models and large-scale foundation models, we introduce **OmniKD**, a unified distillation framework (Section 3.3). OmniKD integrates five complementary signals (logits, features, attention, relational geometry, and contextual similarity) into a single objective, allowing small students to inherit rich supervision without dataset-specific loss engineering or manual loss design.

Table 6 benchmarks OmniSegNet variants against a fine-tuned SAM teacher (ViT-B, 375M parameters) and a wide spectrum of baselines: (i) classical CNNs (U-Net), (ii) light weight compressions (MobileSAM, FastSAM), (iii) transformer-based designs (Swin-UNet, TransUNet, SegFormer), (iv) large auto-configured pipelines (nnU-Net), and (v) recent lightweight foundation-model derivatives (MedSAM-lite, UNesT, Medical SAM Adapter). Despite being up to **38× smaller**, OmniSegNet-Acc (4.7M - 6M) matches SAM within 1–2 pp Dice across six heterogeneous datasets. Balanced (3M - 3.3M) and Compact (1.5M - 1.6M) models consistently outperform size-matched alternatives.

In contrast, heavy transformer or foundation-model variants either fall short in Dice under identical training or demand significantly higher parameters, FLOPs, and latency. These results underscore that OmniSegNet is not only accurate but also uniquely suited for practical

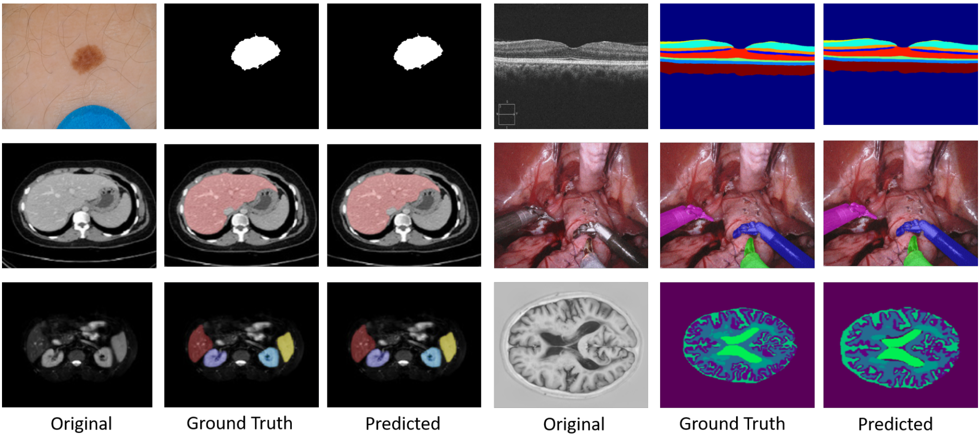

Figure 2: Qualitative results across diverse modalities using our OmniSegNet-B(3.1M) model. **Left column (top to bottom)**: ISIC dermoscopic image with melanoma segmentation, CHAOS-CT with liver delineation, and CHAOS-MRI with multi-organ segmentation. **Right column (top to bottom)**: Retinal OCT with layered retinal structure segmentation, intra-operative video frames with instrument segmentation, and MRBrainS18 with brain tissue segmentation. The model effectively generalizes across domains, capturing both coarse and fine anatomical details.

edge deployment under hospital-grade hardware constraints.

Model (Params)	**ISIC17**	**CHAOS CT**	**CHAOS MRI**	**OCT**	**Intra-op**	**MRBrainS18**
SAM (ViT-B, 375M)	0.97±0.01	0.98±0.01	0.95±0.01	0.97±0.01	0.97±0.01	0.95±0.01
OmniSegNet-A (4.7M-6M)	**0.97±0.01**†	**0.96±0.01**	**0.94±0.01**	**0.97±0.01**†	**0.96±0.01**	**0.94±0.01**
OmniSegNet-B (3M-3.3M)	0.95±0.01	0.93±0.01	0.92±0.01	0.90±0.01	0.91±0.01	0.91±0.01
OmniSegNet-C (1.5M-1.6M)	0.87±0.02	0.92±0.01	0.90±0.02	0.87±0.02	0.80±0.02	0.88±0.02
U-Net (7.5M)	0.84±0.02	0.89±0.02	0.87±0.02	0.85±0.02	0.78±0.02	0.86±0.02
MobileSAM [42] (10M)	0.86±0.01	0.91±0.01	0.89±0.01	0.86±0.01	0.79±0.02	0.88±0.01
FastSAM [14] (10M)	0.86±0.01	0.91±0.01	0.89±0.01	0.86±0.01	0.79±0.02	0.88±0.01
Swin-UNet [6] (24M)	0.86±0.01	0.91±0.01	0.89±0.01	0.87±0.01	0.80±0.01	0.89±0.01
TransUNet [7] (95M)	0.87±0.01	0.92±0.01	0.90±0.01	0.88±0.01	0.81±0.01	0.90±0.01
SegFormer-B3 [38] (47M)	0.85±0.01	0.90±0.01	0.89±0.01	0.87±0.01	0.80±0.01	0.89±0.01
Swin-UNETR [12] (100M)	0.86±0.01	0.91±0.01	0.89±0.01	0.87±0.01	0.80±0.01	0.88±0.01
nnU-Net [15] (>100M)	0.88±0.01	0.92±0.01	0.91±0.01	0.88±0.01	0.82±0.01	0.90±0.01
MedSAM [24] (25M)	0.85±0.01	0.90±0.01	0.88±0.01	0.86±0.01	0.79±0.01	0.87±0.01
UNesT [40] (20M)	0.84±0.01	0.89±0.01	0.87±0.01	0.85±0.01	0.78±0.01	0.86±0.01
Medical SAM Adapter [37] (30M)	0.85±0.01	0.90±0.01	0.88±0.01	0.86±0.01	0.79±0.01	0.87±0.01

Table 6: Mean±std Dice scores across six medical segmentation datasets (3-fold CV). Bold indicates the best mean. Significance was tested using paired two-sided t-tests over CV folds with Benjamini–Hochberg FDR correction (q=0.05) [2]. † denotes results not significantly different from the row-best.

We observed that SAM transfers knowledge unevenly across domains: strong gains in dermoscopy (ISIC) and CT/MRI, but relatively smaller improvements in highly specialized domains like retinal OCT or intra-operative video, where visual statistics diverge strongly from SAM's pretraining corpus. Nevertheless, OmniKD still improves Dice by 3–5pp over non-distilled students in these cases, indicating that even partial transfer of SAM's general representations provides value in specialized domains.

4.4 Ablation Study

We conducted a systematic ablation on OmniSegNet and OmniKD across six diverse medical imaging datasets (ISIC, CHAOS-CT, CHAOS-MRI, Retinal OCT, Intraoperative, and MRBrainS18). Each ablation disables a structural module (SE blocks, ASPP, compound scaling) or distillation component (logits, features, attention, relational, contextual). Results in Table 7 are reported as mean±std Dice over 3 folds; † indicates a statistically significant drop (p<0.05, paired t-test vs. Full OmniSegNet-B). Similar ablations for Compact and Accurate variants are presented in the Supplementary.

Variant	ISIC Dice	CHAOS CT	CHAOS MRI	Retinal OCT	Intra-Op	MRBrainS18
Full OmniSegNet-B	**0.95±0.01**	**0.93±0.01**	**0.92±0.01**	**0.90±0.01**	**0.91±0.01**	**0.91±0.01**
w/o SE Blocks	0.84±0.02†	0.87±0.02†	0.87±0.02†	0.89±0.02†	0.83±0.02†	0.81±0.02†
w/o ASPP Module	0.83±0.02†	0.85±0.02†	0.85±0.02†	0.87±0.02†	0.81±0.02†	0.80±0.02†
No Compound Scaling	0.82±0.02†	0.86±0.02†	0.86±0.02†	0.88±0.02†	0.80±0.02†	0.79±0.02†
Student Only (no KD)	0.80±0.02†	0.84±0.02†	0.84±0.02†	0.86±0.02†	0.78±0.02†	0.75±0.02†
+ Logits KD	0.83±0.02†	0.86±0.02†	0.87±0.02†	0.88±0.02†	0.81±0.02†	0.78±0.02†
+ Features (FitNets)	0.84±0.02†	0.87±0.02†	0.88±0.02†	0.90±0.02†	0.82±0.02†	0.79±0.02†
+ Attention Transfer	0.85±0.02†	0.88±0.02†	0.88±0.02†	0.91±0.02†	0.83±0.02†	0.81±0.02†
+ Relational Distillation	0.86±0.02†	0.88±0.02†	0.89±0.02†	0.91±0.02†	0.84±0.02†	0.82±0.02†
+ Contextual Similarity (final OmniKD)	**0.95±0.01**	**0.93±0.01**	**0.92±0.01**	**0.90±0.01**	**0.91±0.01**	**0.91±0.01**

Table 7: Ablation study on the OmniSegNet-Balanced variant (mean±std Dice over 3-fold CV). † indicates a significant drop compared to the full model (p<0.05, paired t-test with Benjamini–Hochberg correction). Each architectural module and OmniKD component contributes to the overall performance.

5 Conclusion

We presented **OmniSegNet**, a unified framework that combines scalable architecture design with multi-term knowledge distillation to address the diverse challenges of medical image segmentation. By integrating interpretable compound scaling with latency-aware neural architecture search, the framework adapts seamlessly to different deployment scenarios ranging from ultra-low edge devices to hospital-grade systems. Alongside, the proposed **OmniKD** strategy consolidates multiple supervision signals including logits, features, attention, relational, and contextual cues into a single objective, enabling compact models to inherit strong generalization ability from large foundation models like SAM without task-specific tuning.

Through extensive experiments, OmniSegNet demonstrated consistent gains across six diverse datasets, delivering accuracy comparable to heavyweight baselines at a fraction of the computational cost while maintaining real-time efficiency on CPUs and micro-GPUs. This positions the framework as a practical pathway towards efficient, privacy-preserving, and generalizable medical image analytics. Looking forward, we envision extending OmniSegNet under the **Knowledge-Rich, Privacy-Preserving initiative** to advance universal distillation, federated optimization, and interpretability modules, paving the way for trustworthy, edge-ready clinical deployment at scale.

References

[1] Ahmed Al Qurri and Mohamed Almekkawy. Improved unet with attention for medical image segmentation. *Sensors*, 23(20):8589, 2023.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[3] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In *International Conference on Learning Representations (ICLR)*, 2019.

[4] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations (ICLR)*, 2020.

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[6] Hong Chang, Shan You, Xinge Zhang, Xiaohui Wu, and Zheng-Jun Zha. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[7] Jianpeng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017.

[9] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[10] Soma Dasgupta, Swarnava Dey, Avik Ghose, Arijit Mukherjee, and Arpan Pal. Y-net: A compact edge-friendly model for multi-modal medical image segmentation. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 243–248. IEEE Computer Society, 2025.

[11] Swarnava Dey and Soma Dasgupta. A genetic tuner for fixed-skeleton tinyml models. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 556–561. IEEE, 2024.

[12] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin-unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Proceedings of the International MICCAI Brain Lesion Workshop*, pages 272–284. Springer, 2022.

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[15] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-adapting framework for u-net-based medical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[16] Yesmina Jaafra, Jean Luc Laurent, Aline Deruyver, and Mohamed Saber Naceur. Reinforcement learning for neural architecture search: A review. *Image and Vision Computing*, 89:57–66, 2019.

[17] Atilla Ersoy Kavur, Cem Gezer, Mustafa Barış, and et al. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[18] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, and et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[19] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. URL https://arxiv.org/abs/2410.17725.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Laura Rolland, Rachel Gustafson, Tete Xiao, Samuel Whitehead, Alexander Clegg, Alexander Boyko, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[21] HJ Kuijf and E Bennink. Mrbrains18, grand challenge on mr brain segmentation at miccai 2018, 2019.

[22] Yifan Liu, Yunchao Tian, Yongxin Wang, Zhi Zhu, Chao Xu, and Yipeng Lin. Structured knowledge distillation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613, 2019.

[23] Haoran Lu, Yifei She, Jun Tie, and Shengzhou Xu. Half-unet: A simplified u-net architecture for medical image segmentation. *Frontiers in Neuroinformatics*, 16:911679, 2022.

[24] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[25] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 552–568, 2018.

[26] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[27] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, 2019.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[29] Romero A Ballas N Kahou SE. Fitnets: hints for thin deep nets. *Fitnets: hints for thin deep nets*, 2014.

[30] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2019.

[31] Mupparaju Sohan, Thotakura Sai Ram, and Ch Venkata Rami Reddy. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, pages 529–545. Springer, 2024.

[32] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114, 2019.

[33] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820–2828, 2019.

[34] Fenghe Tang, Jianrui Ding, Quan Quan, Lingtao Wang, Chunping Ning, and S Kevin Zhou. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[35] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019.

[36] Yu Weng, Tianbao Zhou, Yujie Li, and Xiaoyu Qiu. Nas-unet: Neural architecture search for medical image segmentation. *IEEE access*, 7:44247–44257, 2019.

[37] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[38] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[39] Xiangxiang Yu, Wenqing Wu, et al. Mobilesam: Mobile-friendly segment anything model with tinyvit. *arXiv preprint arXiv:2306.14289*, 2023.

[40] Xin Yu, Qi Yang, Yinchi Zhou, Leon Y Cai, Riqiang Gao, Ho Hin Lee, Thomas Li, Shunxing Bao, Zhoubing Xu, Thomas A Lasko, et al. Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. *Medical Image Analysis*, 90:102939, 2023.

[41] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017.

[42] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[43] Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579*, 2023.

[44] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[45] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.