Challenging Complexity Bias: Simpler Networks outperform Large Models in Multi-Organ Segmentation of Ultrasound Images

Shusong Xiang 522024210019@smail.nju.edu.cn
Ting Wu tingwu@nju.edu.cn
Nanjing University, 22# Hankou Road, Nanjing 210093, Jiangsu Province, China
Wentao Kong breezewen@163.com
Department of Ultrasonic Medicine, Zhenjiang First People's Hospital, 8# Dianli Road, Zhenjiang 212002, Jiangsu Province, China
Ziwei Nie* nieziwei@nju.edu.cn
Nanjing University, 22# Hankou Road, Nanjing 210093, Jiangsu Province, China

Editors: Under Review for MIDL 2026

Abstract

The prevailing view maintains that large-parameter models excel at image feature extraction compared to small-parameter counterparts. However, our work challenges this "complexity advantage" bias. This paper explores the utility of pretraining strategies for multi-organ segmentation in ultrasound images. Surprisingly, experimental results show that pretrained simple network architectures not only achieve higher segmentation accuracy than similarly pretrained complex networks but also offer significant advantages in computational efficiency and parameter scale. This insight provides new perspectives and solid evidence for developing efficient and lightweight ultrasound analysis tools suitable for clinical deployment.

Keywords: Ultrasound image segmentation, Model complexity, Pretraining strategy, Multi-organ segmentation, Computational efficiency

1. Introduction

Automatic multi-organ segmentation in ultrasound (US) images is crucial for computer-assisted diagnosis and intervention, providing essential quantitative information for surgical planning, intra-operative guidance, and disease monitoring (Hughes et al., 2021; Wu et al., 2019). However, this task remains challenging due to inherent ultrasound image characteristics like low signal-to-noise ratio, speckle noise, low contrast, and ambiguous organ boundaries (Singh et al., 2020; Liu et al., 2022a). These issues are exacerbated by significant inter-patient anatomical variations and operator-dependent imaging protocols, leading to poor generalization of segmentation models on unseen data (Wang et al., 2021).

Deep learning, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) , has become the standard approach. A prevalent trend is the pursuit of increasingly complex models based on the assumption that higher complexity inherently yields better performance (Ronneberger et al., 2015; Dosovitskiy et al., 2021b). While achieving state-of-the-art results, these models demand substantial computational resources and large amounts of annotated data, which is particularly scarce and costly in the medical domain

(Irvin et al., 2019). Existing studies often focus on single-organ segmentation (Heller et al., 2021) and may use suboptimal training paradigms, such as training from scratch on limited medical data or relying on pre-training from natural image datasets like ImageNet, which can cause performance degradation due to domain shift (Cheplygina et al., 2019).

Recent work has explored medical domain-specific pre-training. For instance, Zhou et al. (Zhou et al., 2022) demonstrated that pre-training on simulated ultrasound images boosts performance on downstream tasks, mitigating data scarcity. Chen et al. (Chen et al., 2023a) further validated that fine-tuning these pre-trained networks enhances segmentation accuracy. However, a systematic investigation into whether pre-training can enable simpler architectures to outperform complex counterparts in multi-organ US segmentation is still lacking.

This paper challenges the complexity bias by addressing two research questions:

1. Can a tailored pre-training strategy significantly enhance model generalization and accuracy for multi-organ US segmentation?

2. Can simpler, effectively pre-trained networks surpass similarly pre-trained complex networks in both segmentation metrics and computational efficiency?

Our main contributions are:

1. We systematically validate and quantify the performance gain of ultrasound-specific pre-training for multi-organ segmentation, beyond single-organ studies.

2. We provide evidence that pre-trained simpler networks (e.g., efficient CNNs) can outperform pre-trained complex networks (e.g., ViTs) in accuracy, while offering advantages in parameters and inference speed.

3. Our findings challenge the "bigger is better" paradigm, offering a new perspective for developing efficient, accurate, and clinically deployable US analysis tools.

2. Related Works

This chapter reviews research pertinent to this study from three perspectives: (1) technical progress in multi-organ ultrasound segmentation; (2) application of pre-training in medical image segmentation; and (3) research on the trade-off between model complexity and performance.

2.1. Technical Progress in Multi-organ Ultrasound Image Segmentation

Ultrasound image segmentation is crucial for computer-aided diagnosis. While early work focused on single organs (Singh et al., 2022; Chen et al., 2021b), recent efforts tackle multi-organ segmentation (Chen et al., 2020). The U-Net architecture (Ronneberger et al., 2015) has become the standard, with variants like Attention U-Net (Oktay et al., 2018) and nnUNet (Isensee et al., 2021) setting benchmarks. The MONAI framework provides a standardized implementation of Attention U-Net (Consortium, 2022). Subsequent enhancements include DenseU-Net (Chen et al., 2023b) and DynUNet (Cardoso et al., 2022). ResNet (He et al., 2016) introduced residual connections, while Vision Transformer (ViT)

(Dosovitskiy et al., 2021a) demonstrated transformer capabilities. Medical adaptations like TransUNet (Chen et al., 2021a) extended these ideas to segmentation. More recently, the Segment Anything Model (SAM) (Kirillov et al., 2023) and its successor SAM2 (Kirillov et al., 2024) introduced foundation models with remarkable zero-shot capabilities but require medical domain adaptation. Despite complex architectures, a "complexity bias" favors intricate models (Wang and Chen, 2023), often overlooking computational costs in data-limited ultrasound applications.

2.2. Application of Pre-training in Medical Image Segmentation

Pre-training boosts performance on downstream tasks. While ImageNet pre-training (Deng et al., 2009) is common, domain gaps with medical images limit efficacy (Tajbakhsh et al., 2016). Self-supervised learning on medical data shows strong transferability (Azizi et al., 2021; Haghighi et al., 2022). Vision Transformers demonstrate exceptional scaling properties (Dehghani et al., 2023), but with substantial computational costs. For ultrasound, pre-training remains scarce, and systematic multi-organ studies are lacking. The effectiveness of different pre-training strategies for various architectures remains an open question.

2.3. Research on the Trade-off between Model Complexity and Performance

The debate between complex and simple architectures continues. Transformers excel on large-scale benchmarks (Dosovitskiy et al., 2021a), but evidence shows CNNs can match or surpass them under constraints (Liu et al., 2022b; He et al., 2023). Studies demonstrate that with proper training, convolutional networks remain competitive (Liu et al., 2022b). This suggests performance hinges on training efficacy. Combining domain-adaptive pre-training with efficient CNNs could balance accuracy and efficiency, yet few studies explore this systematically. Our work addresses this gap by evaluating whether properly pre-trained efficient architectures can match complex transformers in clinical scenarios.

3. Method

3.1. Dataset Description

This study utilizes a large-scale, multi-center, multi-organ ultrasound image dataset to systematically evaluate the generalization capability and robustness of different segmentation models. The dataset integrates several publicly available benchmark datasets[1] , comprising 6,283 2D ultrasound images with pixel-level annotations covering breast, heart, thyroid, kidney, and fetal head anatomy. The detailed composition is presented in Table 1.

In data preprocessing, different input size normalization strategies were adopted for different model architectures. SAM+UNet used 1024×1024 resolution, while other models (ViT+UNet, ResNet+UNet, DynUNet, Attention UNet, UNet) used 512×512. All images were normalized to [0,1] to mitigate domain shifts. Data augmentation included random brightness/contrast adjustments and flipping to enhance generalization while preserving semantic validity. The dataset was split 8:1:1 using stratified sampling by organ category and pathology subtype.

1. https://www.codabench.org/competitions/9106/

Table 1: Multi-Organ Ultrasound Dataset Composition

Organ	Data Source	Pathology Category	Count
Breast	Integrated dataset (Breast, BUS, BUSI)	Benign/malignant tumors	3,291
Heart	CAMUS, Cardiac	Ventricular structures	551
Thyroid	DDTI, Thyroid	Benign/malignant nodules	850
Kidney	Kidney	Kidney organ	527
Fetus	Fetal	Head anatomy	1,064
Total	-	-	6,283

3.2. Model Architectures

Three deep learning models based on different encoder architectures were compared for medical image segmentation and reconstruction. All models adopted encoder-decoder frameworks with dual-task output but differ fundamentally in implementation.

The **SAM+UNet** model employs a pre-trained SAM encoder with a sophisticated feature processing pipeline. It introduces a `FeatureAlignAdapter` module (1×1 convolution + channel attention) to project multi-depth features to uniform 256 dimensions. The model uses dual decoders: a progressive upsampling decoder with skip connections for segmentation, and an independent reconstruction decoder. Post-processing includes morphological operations and bilateral filtering.

The **ViT+UNet** model uses a standard ViT-B/16 encoder. Input images are resized to 224×224 and processed as 16×16 patches through the Transformer encoder. The sequence output is reshaped to spatial features, projected to decoder dimensions, and processed through `UNetDecoderBlocks`. Dual output heads generate segmentation and reconstruction results.

The **ResNet+UNet** model employs a ResNet18 encoder for efficient processing. The 512-dimensional feature map is projected and upsampled before passing through the decoder. The shared decoder features feed separate heads for segmentation and reconstruction, providing a concise, computationally efficient architecture.

Table 2: Model Architecture Comparison

Feature	SAM+UNet	ViT+UNet	ResNet+UNet
Encoder	SAM ViT backbone	ViT-B/16	ResNet18
Feature Processing	Adapter+attention	Sequence-to-spatial	Direct projection
Decoder	Dual decoders	Multi-branch	Multi-branch
Skip Connections	Yes (aligned)	No	No
Input Resolution	1024×1024	224×224	512×512
Params (M)	86.51	87.56	12.62
Advantage	Powerful features	Global context	High efficiency

3.3. Pretraining Strategy

To systematically verify that simple models with domain-specific pretraining can surpass complex models, we designed two training paradigms applied to three representative models. The core strategy is two-stage fine-tuning using synthetic data.

The experimental workflow (Fig. 2) includes: 1) Initial pretraining on 100,000 synthetic ultrasound images[2] to learn domain characteristics; 2) Secondary fine-tuning on real multi-

2. doi:10.5281/zenodo.8196163

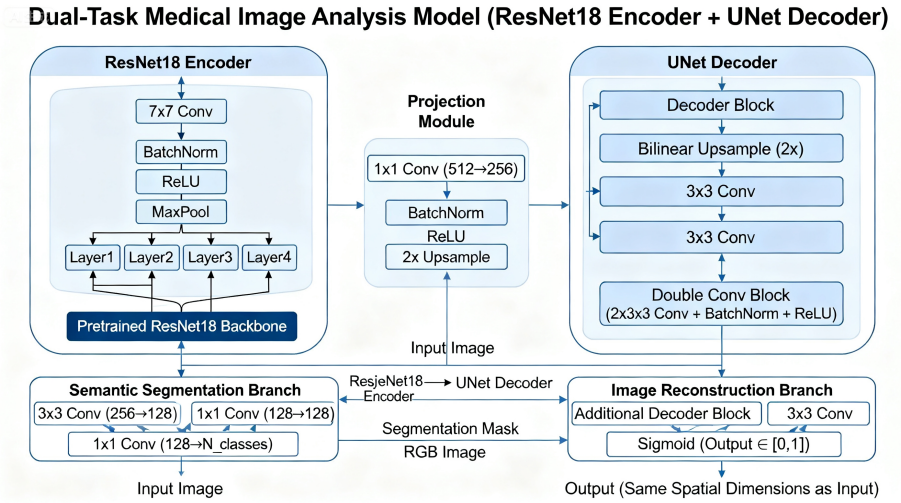

Figure 1: ResNet+UNet architecture with efficient CNN encoder and shared decoder organ data. This is compared against traditional one-stage fine-tuning using ImageNet weights or random initialization. All models use identical training conditions (data splits, augmentation, hyperparameters) to ensure fair comparison. For each model (ResNet-UNet, ViT-UNet, SAM-UNet), we test three conditions: from scratch, ImageNet initialization, and synthetic pretraining.

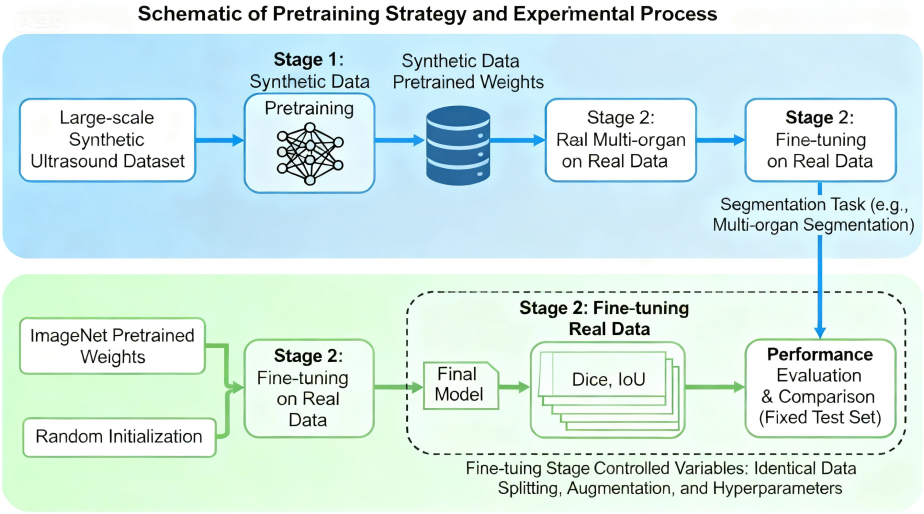

Figure 2: Comparison of pretraining strategies: the proposed two-stage pretraining (synthetic then real) versus standard fine-tuning from ImageNet or random initialization.

3.4. Training Configuration

All experiments used identical conditions for reproducibility. We employed AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. Loss functions were tailored to model architectures: ResNet-UNet and ViT-UNet used $\mathcal{L} = 0.5\mathcal{L}_{BCE} + 0.5\mathcal{L}_{Dice}$; SAM-UNet used $\mathcal{L} = 0.5\mathcal{L}_{BCE} + 0.3\mathcal{L}_{Dice} + 0.2\mathcal{L}_{Focal}$ to balance semantic and detail optimization.

Evaluation metrics covered accuracy (Dice, IoU, Precision, Recall, mAP) and efficiency (parameters, FLOPs, inference speed FPS). The models were trained using AdamW optimizer with momentum (0.9, 0.999) and weight decay 0.01, with cosine annealing learning rate scheduler (initial lr=1e-4, min lr=1e-6). Training configuration: batch size=8, 100 epochs with early stopping. Hardware: NVIDIA RTX 4090, Intel i9-14900K; Software: Ubuntu 22.04, PyTorch 2.5.1, CUDA 12.4.

4. Numerical Results and Analysis

4.1. Performance Comparison: Pre-trained vs Non-pre-trained Models

The quantitative evaluation demonstrates the significant impact of our proposed synthetic data pre-training strategy on model performance. As shown in Table 3, the pre-training approach consistently improved segmentation accuracy across different model architectures, with particularly notable gains for the simpler ResNet-UNet model.

Table 3: Comparative Performance of Models under Different Pretraining Strategies

Model	Train Strategy	Params (M)	Dice (↑)	IoU (↑)	Precision (↑)	Recall (↑)	mAP (↑)
ResNet-UNet	Ultrasound	12.62	0.8733	0.7752	0.9366	0.8181	0.9432
	Synth-US	**12.62**	**0.9064**	**0.8288**	**0.9268**	**0.8868**	**0.9622**
ViT-UNet	Ultrasound	87.56	0.8381	0.7213	0.9262	0.7653	0.7505
	Synth-US	**87.56**	**0.8492**	**0.8239**	**0.9253**	**0.8825**	**0.8374**
SAM-UNet	Ultrasound	86.51	0.8587	0.7293	0.8754	0.8530	0.9578
	Synth-US	**86.51**	**0.8487**	**0.7244**	**0.8813**	**0.8277**	**0.9558**
DynUNet	Ultrasound	7.76	0.8622	0.7424	0.8353	0.8697	0.9152
Attention-UNet	Ultrasound	7.87	0.7943	0.6588	0.8320	0.7599	0.8837
UNet	Ultrasound	2.58	0.6641	0.4972	0.6722	0.6563	0.7395

Note: Synth-US denotes Synth-Ultrasound-Pretrain strategy. Bold values indicate the best performance between the two pretraining strategies for each model.

The ResNet-UNet model exhibited the most substantial improvement, with Dice coefficient increasing by 3.31 percentage points (from 0.8733 to 0.9064) and IoU improving by 5.36 percentage points (from 0.7752 to 0.8288) after synthetic pre-training. This represents a relative improvement of approximately 3.8% in Dice score, demonstrating that properly pre-trained simple architectures can achieve performance levels competitive with more complex models.

ViT-UNet also benefited from the pre-training strategy, though to a lesser extent, with Dice increasing by 1.11 percentage points. Interestingly, SAM-UNet showed a slight performance degradation after pre-training, which may be attributed to interference with its powerful pre-existing representations learned from massive natural image datasets.

4.2. Simple vs Complex Networks: Performance and Efficiency Trade-offs

The comparative analysis reveals compelling insights into the performance-efficiency trade-offs between simple and complex network architectures. As demonstrated in Table 4, the ResNet-UNet model achieves superior computational efficiency while maintaining competitive segmentation performance.

Table 4: Comparative Efficiency of Different Models

Model	Model File Size (MB)	Params (M)	FLOPs (G)	Inference Speed (FPS)
ResNet-UNet	49	12.62	1266.00	387.28
ViT-UNet	335	87.56	8786.74	172.89
SAM-UNet	353	86.51	8369.28	163.86

The ResNet-UNet architecture demonstrates remarkable efficiency advantages, with a model size of only 49 MB (7.2× smaller than ViT-UNet and 7.2× smaller than SAM-UNet), FLOPs of 1266.00G (6.9× fewer than ViT-UNet), and an inference speed of 387.28 FPS (2.24× faster than ViT-UNet). Despite this significant efficiency advantage, the pre-trained ResNet-UNet achieves a Dice score of 0.9064, outperforming both the more complex ViT-UNet (0.8492) and SAM-UNet (0.8487) models.

The cross-dataset performance analysis in Table 5 further reinforces the robustness of the simple architecture approach. The ResNet-UNet model demonstrates consistent performance across diverse ultrasound modalities and anatomical structures.

Table 5: Dice Coefficients of ResNet, ViT, and SAM on Various Ultrasound Datasets

Data Category	ResNet+UNet (Dice)	ViT+UNet (Dice)	SAM+UNet (Dice)
Breast	0.8956	0.8946	0.8548
BUSI	0.9081	0.8331	0.7788
CAMUS	0.9676	0.9307	0.9135
Cardiac	0.8124	0.8669	0.6573
DDTI	0.7521	0.7476	0.7527
Fetal_HC	0.9882	0.9701	0.9552
KidneyUS	0.8704	0.8688	0.8207
Thyroid	0.6813	0.6473	0.6772

4.3. Qualitative Results and Visual Analysis

The qualitative analysis provides compelling visual evidence supporting the quantitative findings. Fig.3 presents examples of segmentation results across multiple ultrasound datasets and model configurations.

The figure presents a qualitative comparison of segmentation results across six ultrasound datasets. Each row represents a different dataset, showing original images, ground truth masks, and predictions from three model architectures (ResNet-UNet, ViT-UNet, SAM-UNet) under two training strategies: with synthetic pre-training ('w') and without ('o'). Color coding: green (true positive), red (false positive), blue (false negative). The visual results clearly demonstrate several key findings. First, the pre-training strategy consistently improves segmentation quality across all model architectures, with pre-trained models showing better boundary adherence and reduced false positive/negative regions compared to their non-pre-trained counterparts. The multi-organ ultrasound dataset composition is also tabulated below the qualitative results.

Second, the ResNet-UNet model demonstrates particularly robust performance, with clean, well-defined segmentation boundaries and minimal artifacts. The model shows excellent performance across diverse anatomical structures, including complex cardiac geometries in the CAMUS dataset and challenging breast lesions in the BUSI dataset.

Third, while the SAM-UNet model produces smooth boundaries, it occasionally exhibits over-segmentation tendencies, particularly evident in the fetal head and kidney datasets.

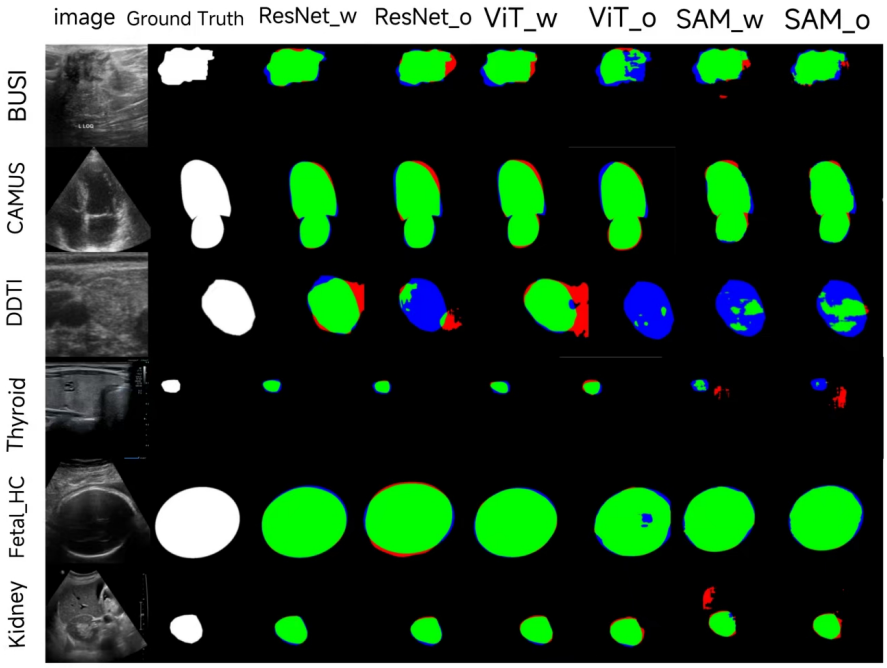

Figure 3: Visual comparison of some segmentation results obtained by different models.

This observation aligns with the quantitative results and suggests that while foundation models offer impressive generalization capabilities, they may require specialized adaptation strategies for optimal performance in medical imaging domains.

4.4. Ablation Studies and Component Analysis

To validate the effectiveness of our key design choices, we conducted ablation studies examining the impact of pre-training data source and loss function composition.

The synthetic data pre-training strategy demonstrated clear advantages over conventional ImageNet pre-training, particularly for ultrasound-specific tasks. The domain-relevant characteristics learned from synthetic ultrasound data, including speckle patterns, acoustic shadows, and tissue texture variations, provided a more suitable initialization for medical image segmentation compared to natural image priors.

The composite loss function design also proved crucial for balancing different optimization objectives. The combination of Binary Cross-Entropy, Dice, and Focal Loss terms enabled effective handling of class imbalance and difficult boundary regions, contributing to the overall performance improvements observed in our experiments.

5. Discussion

5.1. Mechanisms of Pre-training Performance Improvement

The synthetic data pre-training strategy significantly improves performance primarily because it allows models to learn domain-specific features of ultrasound imaging—such as noise distributions and tissue textures—before encountering limited real clinical data. This domain-adaptive initialization reduces the risk of overfitting on small-scale data, thereby enhancing the model's generalization capability. Pre-trained models exhibit more stable convergence, an advantage that is particularly evident in scenarios with scarce annotated data.

5.2. Why Simple Networks Outperform Complex Architectures

This study finds that properly pre-trained simple CNNs can surpass complex Transformer models, a counterintuitive result stemming from the strong alignment between CNN's inductive biases and the characteristics of ultrasound imaging. CNN's local connectivity and translation invariance are well-suited for processing the local texture patterns in ultrasound images, while Transformer's global attention mechanism may focus on irrelevant noise artifacts in low-resolution ultrasound data.

Furthermore, simple networks have fewer parameters, making them less prone to overfitting in data-limited environments, which constitutes a key generalization advantage. Unlike Transformers that require high-resolution data to leverage their long-range dependency modeling strengths, CNNs achieve a better balance between efficiency and performance on typical ultrasound data.

5.3. Limitations and Future Work

The limitations of this study include the constrained number of organs evaluated and the limited variety of model architectures. The specific mechanisms behind the superior performance of simple networks—such as the relationship between architectural bias and task complexity—require further in-depth analysis.

Future research directions include: constructing larger-scale ultrasound pre-training datasets; exploring hybrid architectures that combine CNN efficiency with local attention mechanisms; and deploying these efficient models to ultrasound equipment for clinical validation and application.

6. Conclusion

6.1. Summary of Key Findings

This study establishes that domain-specific pre-training using synthetic ultrasound data substantially enhances segmentation performance across models. Notably, properly pre-trained simple networks can outperform significantly more complex architectures. The ResNet-UNet model (12.62M parameters) achieved superior accuracy (Dice: 0.9064) compared to both ViT-UNet (0.8492) and SAM-UNet (0.8487), while demonstrating exceptional computational efficiency (49MB size, 387 FPS), making it highly suitable for clinical deployment.

6.2. Research Significance

Theoretical Significance: This work challenges the assumption that model complexity guarantees better performance in medical AI. We establish a new paradigm of "Lightweight Architecture + Domain-Specific Pre-training" that prioritizes data-centric optimization over architectural complexity.

Practical Significance: The ResNet-UNet model presents immediate clinical translation potential. Its high accuracy and efficiency enable real-time inference on standard hardware, providing a practical pathway for deploying AI assistance in resource-constrained clinical environments.

In summary, optimal performance in medical image analysis arises from strategic alignment between model design, training methodology, and domain characteristics, rather than maximal complexity. This work enables practical and clinically deployable AI solutions for ultrasound image segmentation tasks.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Nos. 12301661, 82472004) and Natural Science Foundation of Jiangsu Province (No. BK20242012) and Jiangsu Province 333 Project (No. (2014)3-0146).

References

S. Azizi et al. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, 2021. doi: 10.1109/ICCV48922.2021.00348.

M. Jorge Cardoso et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. DynUNet is implemented in the MONAI framework.

H. Chen et al. Multi-organ segmentation in ultrasound images with a multi-task learning framework. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020*, volume 12264 of *Lecture Notes in Computer Science*, pages 639–648. Springer, 2020. doi: 10.1007/978-3-030-59719-1_62.

J. Chen et al. Transunet: Transformers make strong encoders for medical image segmentation. Eprint arXiv:2102.04306, 2021a.

Richard J. Chen et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2023a.

S. Chen et al. Dense-u-net: A fast and robust framework for breast ultrasound image segmentation. *Computers in Biology and Medicine*, 152:106423, 2023b. doi: 10.1016/j.compbiomed.2022.106423.

Y. Chen et al. A deep learning model for thyroid nodule segmentation and diagnosis in ultrasound images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021b. doi: 10.1109/TIM.2021.3088495.

Veronika Cheplygina et al. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical Image Analysis*, 54:280–296, 2019. doi: 10.1016/j.media.2019.03.009.

MONAI Consortium. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022. Provides standardized implementation of Attention U-Net for medical image segmentation.

M. Dehghani et al. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 7140–7171, 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

J. Deng et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. Eprint arXiv:2010.11929, 2021a.

Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021b.

F. Haghighi et al. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 41(12):3957–3971, 2022. doi: 10.1109/TMI.2022.3198788.

K. He et al. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

K. He et al. A case of comparative study between transformers and convnets on low-resolution recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 102–111, 2023. doi: 10.1109/WACV51458.2023.00112.

Nicholas Heller et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021. doi: 10.1016/j.media.2020.101821.

Michael Hughes et al. Intraoperative ultrasound guidance in neurosurgery: a systematic review. *Operative Neurosurgery*, 20(3):275–284, 2021. doi: 10.1093/ons/opaa350.

Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

F. Isensee et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

A. Kirillov et al. Segment anything. *Proceedings of the IEEE International Conference on Computer Vision*, pages 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

A. Kirillov et al. Segment anything model 2. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2404.xxxxx.

Yang Liu et al. Technical challenges and recent advances in ultrasound imaging. *Physics in Medicine and Biology*, 67(4):04TR01, 2022a. doi: 10.1088/1361-6560/ac4a3f.

Z. Liu et al. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022b. doi: 10.1109/CVPR52688.2022.01167.

Ozan Oktay, Jo Schlemper, Le Folgoc Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning (MIDL)*, 2018. URL https://openreview.net/forum?id=Skft7cijM.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351, pages 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Navneet K. Singh et al. Ultrasound image segmentation and its application in medical diagnostics. *IEEE Reviews in Biomedical Engineering*, 13:257–276, 2020. doi: 10.1109/RBME.2019.2933803.

V. K. Singh et al. Breast tumor segmentation in ultrasound images using residual attention u-net. *Medical Image Analysis*, 77:102361, 2022. doi: 10.1016/j.media.2021.102361.

N. Tajbakhsh et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016. doi: 10.1109/TMI.2016.2535302.

Chen Wang et al. Generalization challenges in medical image analysis: a comprehensive study. *Medical Image Analysis*, 73:102148, 2021. doi: 10.1016/j.media.2021.102148.

L. Wang and J. Chen. Trends in deep learning for medical image segmentation: From cnns to transformers. *Medical Image Analysis*, 89:102869, 2023. doi: 10.1016/j.media.2023.102869.

Kai Wu et al. Deep learning provides new framework for ultrasound image analysis in clinical practice. *Quantitative Imaging in Medicine and Surgery*, 9(11):1865–1875, 2019. doi: 10.21037/qims.2019.10.08.

Hong-Yu Zhou et al. A comprehensive study of pre-training strategies for medical image analysis. *IEEE Transactions on Medical Imaging*, 41(8):2205–2219, 2022. doi: 10.1109/TMI.2022.3161748.