EfficientMedNeXt: Multi-Receptive Dilated Convolutions for Medical Image Segmentation

Md Mostafijur Rahman, Mustafa Munir, and Radu Marculescu

The University of Texas at Austin, Austin, Texas 78703, USA
mostafijur.rahman@utexas.edu

Abstract. In this work, we introduce EfficientMedNeXt—a lightweight, high-performance segmentation architecture developed through a two-phase optimization process applied to the MedNeXt architecture. To this end, we first optimize the decoder by reducing the high-resolution redundancy and unifying the decoder channels across stages for improved efficiency. Then, we introduce a new Dilated Multi-Receptive Field Block (DMRFB) to capture the multi-scale spatial context efficiently without increasing the kernel sizes and relying on the channel expansion convolutions. Extensive evaluations on BTCV, FeTA, and MSD show that EfficientMedNeXt-L achieves 87.0% DICE score on BTCV (+1.04% over MedNeXt-L) with 96.5% fewer parameters and 77.03% lower FLOPs. In addition, EfficientMedNeXt-S offers comparable DICE score, improved HD95, and 78.1% higher throughput while reducing parameters by 98.5% and FLOPs by 95%. These results demonstrate EfficientMedNeXt's efficiency and accuracy, making it well-suited for real-world clinical applications. Our implementation is available at https://github.com/SLDGroup/EfficientMedNeXt.

Keywords: Medical Image Segmentation · Multi-receptive Convolutions · Dilated Convolutions · Efficient 3D CNN.

1 Introduction

Medical image segmentation plays a crucial role in clinical diagnostics by enabling precise delineation of anatomical structures such as organs, tumors, and lesions. Early deep learning-based segmentation architectures focus primarily on 2D CNNs, with UNet [23] introducing an encoder-decoder paradigm that retains spatial information through skip connections. This architecture inspired various refinements, including UNet++ [31] and AttnUNet [14], which incorporate dense connectivity and attention mechanisms to improve feature extraction. To address the limited receptive fields of CNNs, Transformer-based 2D architectures such as SwinUNet [2], MedT [26], and MISSFormer [9] use self-attention to capture long-range dependencies. Hybrid approaches such as TransUNet [3], EMCAD [22], CASCADE [18], G-CASCADE [20], and MERIT [19] combine Transformer-based encoders with CNN- or GNN-based decoders to balance the

local and global feature extraction. However, despite these advances, all 2D segmentation models inherently suffer from volumetric inconsistency, as they process slices independently and ignore the inter-slice relationships.

To overcome these limitations, 3D segmentation architectures analyze entire volumetric scans holistically using one of three approaches: fully Transformer-based methods, hybrid Transformer-CNN designs, or architectures based solely on convolutional networks. nnUNet [10] automates training configurations based on the dataset characteristics. UNETR [6] integrates Vision Transformers (ViTs) [4] encoder for long-range dependency modeling with a conventional CNN decoder. Hybrid Transformer-CNN architectures such as SwinUNETR [5] and SwinUNETRv2 [7] apply shifted window attention to the encoder and 3D residual convolutions to the decoder. TransBTS [27] combines self-attention encoding with convolutional decoding, while UNeSt [29] enhances local spatial communication via hierarchical patch aggregation. Despite their performance gains, these models still demand substantial GPU resources for inference.

Fully convolutional models like MedNeXt [24] and 3D UX-Net [12] replicate large receptive fields using large-kernel depthwise convolutions. However, MedNeXt and 3D UX-Net rely on fixed single-scale receptive fields, inefficient convolutional block design, redundant high-resolution decoder layers, and excessive channels in the decoder, thus leading to excessive computational cost and memory consumption with suboptimal performance.

To address efficiency, recent models target lightweight designs. UNETR++ [25] introduces paired attention for efficiency, while SegFormer3D [17] uses an all-MLP decoder to reduce computations. SlimUNETR [15] prunes channels and attention blocks to improve speed and reduce memory footprint. EffiDec3D [21] further optimizes architecture by eliminating high-resolution decoder stages and unifying the reduced decoder channel counts across stages.

Despite advances in both heavy and lightweight models, few architectures strike an optimal balance between segmentation accuracy and real-time efficiency. Heavy approaches deliver strong performance, but are impractical for deployment due to high computational and memory demands, while lightweight variants often sacrifice accuracy. An ideal model, therefore, should seamlessly integrate multi-scale spatial information, capturing both global context and local detail, without relying on large kernels or self-attention. It should eliminate redundant computations in the decoder, reducing memory footprint and inference cost, and be architecturally lean, thus enabling real-time processing on clinical hardware. Crucially, this requires a carefully designed core convolutional block that balances receptive field flexibility and computational efficiency.

We introduce EfficientMedNeXt, a segmentation architecture designed to overcome these challenges. Developed through a two-phase architectural optimization of MedNeXt [24], EfficientMedNeXt delivers high performance while dramatically reducing computational complexity. Our design enables effective multi-scale spatial learning while maintaining computational efficiency. Our main contributions are as follows:

1. **New Efficient Architecture:** We introduce EfficientMedNeXt, a new segmentation architecture that effectively balances efficiency and segmentation performance by leveraging architectural optimizations that reduce computational cost while maintaining strong feature representation.
2. **Dilated Multi-Receptive Field Block:** We introduce a new convolutional block that replaces the traditional depthwise convolutions with *multi-receptive dilated depthwise convolutions* and *removes channel expansion convolutions*, thus allowing for efficient receptive field expansion without increasing kernel sizes. Using multi-receptive dilation-based spatial aggregation, our model adaptively captures both smaller and larger contextual features, thus improving segmentation accuracy while keeping computational costs low.
3. **Two-Phase Architecture Optimization:** We present a two-phase architecture optimization strategy that first reduces the MedNeXt's decoder complexity by unifying decoder channels across all stages and removing redundant high-resolution layers, leading to a 72.7% parameters and 53.9% FLOPs reduction with only a minor (-0.38%) DICE score drop. In the second phase, we restore the DICE score by +1.42% with additional computational savings through adaptive multi-scale receptive field aggregation, ensuring an optimal balance between computational efficiency and segmentation performance.

The rest of this paper is organized as follows. Section 2 details our architecture, EfficientMedNeXt. Section 3 describes the experimental setup. Section 4 presents our experimental results. Section 5 summarizes our contributions.

2 Method

2.1 Dilated Multi-Receptive Field Block (DMRFB)

The existing MedNeXt block [24] (Fig. 1c) utilizes a large-kernel depthwise convolution (DWC) followed by a 1×1 convolution for channel expansion with a scaling factor R, before projecting to the output channels. This design captures the single-scale spatial context but relies on channel expansion convolutions, which contribute significantly to computational costs.

To address these inefficiencies, we introduce the **Dilated Multi-Receptive Field Block (DMRFB)**, which *removes the expansion convolution* while *enhancing receptive field diversity* through parallel dilated depthwise convolutions. As shown in Fig. 1b, DMRFB replaces the single large-kernel DWC with three parallel depthwise convolutions: (1) a 1×1 convolution for local feature extraction, (2) a 3×3 depthwise convolution with dilation $D = 1$ for mid-range receptive fields, and (3) a 3×3 depthwise convolution with dilation $D = 2$ for larger spatial contexts. The outputs from these branches are concatenated and processed by the Group Normalization (GN) [28] as in Eq. 1:

$$Y = \sigma\left(\mathrm{GN}\left(\mathrm{Concat}\big(\mathrm{DWC}_{1\times1,S}(X), \mathrm{DWC}_{3\times3,D=1,S}(X), \mathrm{DWC}_{3\times3,D=2,S}(X)\big)\right)\right) \quad (1)$$

$$\mathrm{DMRFB}(X) = \mathrm{Conv}_{1\times1}(Y) + \begin{cases} X, & \text{if } C_{\text{in}} = C_{\text{out}}, \\ \mathrm{Conv}_{1\times1}(X), & \text{otherwise.} \end{cases} \quad (2)$$

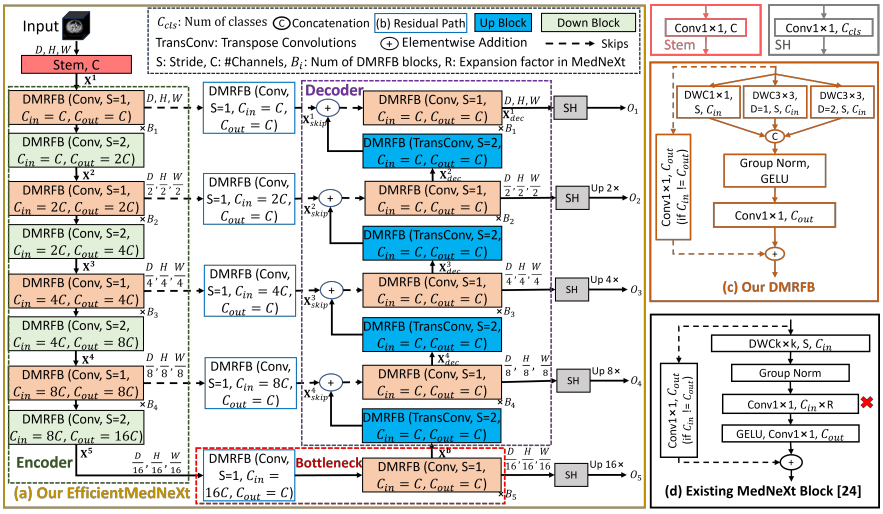

Fig. 1. Our new architecture and components. (a) Our U-shaped EfficientMedNeXt architecture, (b) Our new Dilated Multi-receptive Field Block (DMRFB), and (c) Existing MedNeXt block [24]. Our new Residual Path, down convolution block (Down Block), and up convolution block (Up Block) are based on our DMRFB block. The layer marked ✗ is not used in our DMRFB, thus reducing computational cost. O_1, O_2, O_3, O_4, and O_5 are output segmentation maps from different stages of our network.

where $\sigma(\cdot)$ represents the GELU [8] activation and S is the stride. Afterward, a 1×1 convolution projects the features onto the output channels C_{out}. A final 1×1 convolution is used in residual connection if $C_{in} \neq C_{out}$ as in Eq. 2.

Compared to the MedNeXt block [24], our DMRFB introduces two key optimizations: (1) *removing expansion convolutions*, thus reducing #FLOPs and #Params and (2) *multi-receptive convolutions*, thus capturing multi-scale spatial context without increasing kernel sizes. These modifications enable EfficientMedNeXt to achieve higher segmentation performance with fewer computations.

2.2 EfficientMedNeXt: New Efficient Encoder-Decoder Architecture

As shown in Fig. 1a, EfficientMedNeXt follows an encoder-decoder structure, incorporating a stem layer, an encoder with downsampling, a bottleneck, a decoder with upsampling, and residual paths for skip connection refinement.

Stem Layer. An initial stem layer applies a 1×1 convolution to project the input image $\mathbf{I} \in \mathbb{R}^{C_{\text{in}} \times H \times W \times D}$ to a base feature map with C channels.

Encoder. Each encoder stage $l \in \{1, 2, 3, 4\}$ consists of B_l stacked DMRFB blocks followed by downsampling using DMRFB *with stride 2*, which integrates strided depthwise convolutions to reduce spatial resolution while preserving receptive field expansion. Given an input feature map $\mathbf{X}^l \in \mathbb{R}^{C_l \times H_l \times W_l \times D_l}$ at stage l, the encoder operations are defined as $\mathbf{X}^{l+1} = DMRFB_{S=2}(DMRFB^{B_l}(\mathbf{X}^l))$, where $DMRFB_{S=2}$ applies stride-2 depthwise convolutions for downsampling.

Residual Path for Skip Feature Refinement and Uniform Channel Control. Instead of directly passing the encoder features to the decoder via skip connections,

we refine them using a DMRFB block. This serves two key objectives: (1) enhancing feature representation and (2) ensuring that all skip connection features have *uniform channel dimensions* before aggregation in the decoder. The refined skip features $\mathbf{X}^l_{skip}$ at each stage l are computed as $\mathbf{X}^l_{skip} = DMRFB(\mathbf{X}^l, C^l_{dec})$, where C^l_{dec} ensures that the number of channels in the refined skip features matches the uniform reduced channel dimension of the decoder.

Bottleneck. Before feeding the encoded high-level features ($\mathbf{X}^5$) to the decoder, a bottleneck stage adjusts the number of channels to the decoder's reduced uniform channels using a DMRFB block followed by a series of B_5 stacked DMRFB blocks for feature refinement as $\mathbf{X}^b = DMRFB^{B_5}(DMRFB(\mathbf{X}^5, C^5_{dec}))$.

Decoder. The decoder takes the refined features by the bottleneck layer. Then each decoder stage $l \in \{4, 3, 2, 1\}$ reconstructs feature maps progressively, using DMRFB *with stride 2 and transposed depthwise convolutions* for upsampling, followed by B_l stacked DMRFB blocks for refinement. Given an input feature map from the previous stage $\mathbf{X}^{l+1}_{dec} \in \mathbb{R}^{C_{l+1} \times H_{l+1} \times W_{l+1} \times D_{l+1}}$, the upsampling and refinement process is defined as $\mathbf{X}^l_{dec} = DMRFB^{B_l}(\mathbf{X}^l_{skip} + DMRFB^{\text{trans}}_{S=2}(\mathbf{X}^{l+1}_{dec}))$, where $DMRFB^{\text{trans}}_{S=2}$ applies stride-2 transposed dilated depthwise convolutions.

Multi-Resolution Deep Supervision. To improve optimization and gradient flow, we apply deep supervision by generating segmentation outputs at all five decoder stages using SH. The total loss is calculated as $\mathcal{L}_{\text{total}} = \sum_{i=0}^{4} \lambda_i \mathcal{L}(O_i, G)$, where O_i represents the segmentation prediction from decoder stage i, G is the groundtruth, and $\lambda_i = 1$ is the supervision weight at each stage.

Segmentation Output Selection. During inference, we optimize computational efficiency by selecting the segmentation output from either the final decoder stage (O_1) or the second-to-last decoder stage (O_2), depending on the removal of the high-resolution layer ($H \times W \times D$). This balances accuracy and computational cost, thus making EfficientMedNeXt adaptable to various deployment scenarios.

2.3 Decoder Optimization and Global Network Scaling

Decoder Optimization. Conventional decoders suffer from redundant computations due to excessive channel dimensions and high-resolution processing [21]. Inspired by EffiDec3D [21], we use uniform decoder channels (UDC), aligning all decoder channels with the lowest encoder stage to maintain uniformity as $C^l_{\text{dec}} = \min(\alpha C_{\min}, C_{\max})$, where $C_{\min} = \min(C^1_{\text{enc}}, \ldots, C^5_{\text{enc}})$ ensures minimal redundancy, $\alpha = 1$ scales decoder width, and $C_{\max}$ prevents excessive growth. In addition, we employ high-resolution decoder stage removal (HRR) of ($H \times W \times D$) due to having higher computational overhead with minimal performance gain, thus significantly reducing computation while preserving segmentation accuracy.

Base Channel Scaling (BCS). We also scale the global base number of channels in our EfficientMedNeXt network, starting from the *stem layer*. The lower values of base #channels (16) optimize efficiency, and the larger values (32) improve the representation ability. However, the performance gain with larger base #channels is not always proportional to the computational overhead. Therefore, we introduce two network variants with 16 and 32 base #channels.

3 Experimental Setup

3.1 Datasets

To evaluate the performance of EfficientMedNeXt, we conduct experiments on multiple medical segmentation datasets. **FeTA** [16] consists of 80 T2-weighted infant brain MRIs with annotations for seven distinct tissue types. We randomly split the dataset into 64 training, 8 validation, and 8 testing scans. **BTCV** [11] includes 30 abdominal CT scans with annotations for 13 organs. Following [3], we train on 18 scans, validate on 12, and perform 8-organ and 13-organ segmentation. From **MSD** [1], we use Brain Tumor, Heart, and Lung segmentation tasks, applying an 80:20 train-validation split. See EffiDec3D [21] for more details.

3.2 Implementation Details

Our EfficientMedNeXt is implemented using PyTorch and MONAI (https://monai.io/), with experiments conducted on NVIDIA RTX 6000 (Ada) GPUs with 48GB memory. The input patch size is $96 \times 96 \times 96$, with two cropped sub-volumes for all datasets, except for BTCV. We adopt preprocessing and data augmentation from 3D UX-Net [12] and EffiDec3D [21], including random cropping, flipping, rotation, and intensity normalization.

For loss calculation, we use a combination of DICE and cross-entropy losses. Deep supervision is applied across five decoder stages, and during inference, the output is selected based on network variant for an optimal trade-off between accuracy and efficiency. *Softmax* activation with *Argmax* is applied to decoder outputs, except for Task01_BrainTumor, which uses *Sigmoid* with a 0.5 threshold for multi-level segmentation. Segmentation performance is primarily evaluated using the DICE score, with the 95% Hausdorff Distance (HD95) additionally reported for BTCV 8-organ segmentation.

Training uses the AdamW optimizer [13] with a weight decay of 0.08 and a base learning rate 1×10^{-3} for all datasets. Models are trained for 45,000 steps, except Task01_BrainTumor, which undergoes 60,000 steps. The best model is selected based on validation DICE scores.

4 Results

4.1 Architecture Optimization and Component Ablation

Table 1 analyzes the architectural refinements of MedNeXt to balance precision and efficiency. Starting from baseline MedNeXt-L_K5 (62.99M #Params, 251.09G FLOPs), we systematically cut down computational overhead: our HRR reduces FLOPs by 21% (197.62G) with <0.3% DICE loss, UDC slashes parameters by 72.7% (17.19M) and FLOPs by 53.9% (115.84G), doubling throughput, while DMRFB boosts accuracy (86.06% DICE, 4.62 HD95) via multiple receptive fields. The final BCS scaling (base channels=16) yields EfficientMedNeXt-S: 98.5% fewer #Params (0.92M) and 95% fewer FLOPs (12.46G) vs. baseline, with 85.84% DICE and a 78.1% throughput gain (50.26/s) at -0.12% DICE loss.

Configurations	BCS	HRR	UDC	DMRFB	#Params↓	#FLOPs↓	Thrgh. (/s)↑	%DICE↑	HD95↓
MedNeXt-L_K5	32	No	No	No	62.99M	251.09	11.01	85.96	5.48
+ HRR	32	Yes	No	No	62.93M	197.62G	20.99	85.67	5.59
+ UDC	32	Yes	Yes	No	17.19M	115.84G	26.35	85.58	5.61
+ Our DMRFB	32	Yes	Yes	Yes	3.31M	38.35G	27.94	**86.06**	**4.62**
+ BCS	16	Yes	Yes	Yes	**(-98.5%)0.92M**	**(-95%)12.46G**	**(+78.1%)50.26**	85.84	5.34

Table 1. Results of architecture optimization and new components on BTCV 8-organ segmentation. FLOPs are measured for a 96^3 one-channel input with 9 output classes. Applying Base Channel Scaling (BCS) results in our optimized architecture termed EfficientMedNeXt-S. The computational improvements are shown compared to baseline MedNeXt-L_K5 [24]. Note: High Resolution decoder stage Removal (HRR), Uniform Decoder Channels (UDC), Throughput (Thrgh).

Network Variants	#DMRFB (B_i)	HRR	BCS	#Params	#FLOPs	Thrgh. (/s)	DICE (%)	HD95
EfficientMedNeXt-T	[2,2,2,2,2,2,2,2,2]	Yes	16	0.43M	7.09G	80.00	84.51	7.97
EfficientMedNeXt-S	[3,4,8,8,8,8,8,4,3]	Yes	16	0.92M	12.46G	50.26	85.84	5.34
EfficientMedNeXt-M	[3,4,4,4,4,4,4,4,3]	Yes	32	2.17M	33.34G	29.46	86.59	6.98
EfficientMedNeXt-L	[3,4,4,4,4,4,4,4,3]	No	32	2.19M	57.68G	13.58	**87.00**	**4.45**

Table 2. Performance comparison of EfficientMedNeXt variants across architectural configurations (#DMRFB blocks (B_i), High-Resolution Removal (HRR), and Base Channel Scaling (BCS)) on BTCV 8-organ segmentation. Note: Throughput (Thrgh).

4.2 Architecture Variants Ablation

The architectural and performance trade-offs of the EfficientMedNeXt variants are summarized in Table 2. As shown, the small variant (EfficientMedNeXt-S) strikes a balance, achieving competitive accuracy (DICE: 85.84%, HD95: 5.34) with moderate FLOPs (12.46G) and throughput (50.26/s). Scaling to EfficientMedNeXt-L with HRR and BCS=32 improves DICE to 87.00% and reduces HD95 to 4.45 with a higher #Params (2.19M) and #FLOPs (57.68G).

4.3 Segmentation Results Comparison with SOTA Methods

Table 3 shows that EfficientMedNeXt-L achieves state-of-the-art (SOTA) DICE scores on five datasets (BTCV13, BTCV8, BraT, Lung, Heart), outperforming SwinUNETRv2 (+2.03% BTCV13), MedNeXt-L_K5 (+3.1% Lung), and nnFormer (+7.52% Lung) with at least 28.8× fewer #Params (2.19M vs. 62.99M). EfficientMedNeXt-S surpasses UNETR++ on BTCV8 (+3.44%), BraT (+1.35%), and FeTA (+0.75%) using 46.3× fewer #Params. EfficientMedNeXt-T dominates ultra-light models, achieving 77.15% Lung DICE (+22.08% SegFormer3D, +9.49% SlimUNETR) with only 0.43M parameters (10.5× lighter than SegFormer3D). These results establish a new Pareto frontier, unifying clinical-grade precision (83.29 – 92.98% DICE) with deployability (0.43 – 2.19M #Params, 7.09 – 57.68G FLOPs), resolving the efficiency-accuracy trade-off in medical imaging.

4.4 Qualitative Results Comparison with SOTA Methods

Fig. 2 shows the superior anatomical fidelity of EfficientMedNeXt-L against leading methods on two representative BTCV CT slices. Fig. 2a highlights incorrect

Architecture	Params	FLOPs	Inf. Time	Mem.	Thrgh.	Avg. DICE (%)↑					
	(M)↓	(G)↓	(ms)↓	(GB)↓	(/s)↑	BTCV13	BTCV8	FeTA	BraT	Lung	Heart
UNETR [6]	92.78	82.70	11.56	0.77	86.49	74.96	81.97	84.19	75.69	65.38	91.42
nnFormer [30]	149.32	273.41	20.17	0.94	49.58	78.28	84.56	87.03	74.79	69.79	92.21
TransBTS [27]	31.58	110.66	14.49	0.51	69.01	78.87	83.67	87.18	77.82	63.57	90.12
SwinUNETR [5]	62.19	329.20	47.98	1.51	20.84	80.13	84.42	86.71	79.06	65.12	91.92
SwinUNETRv2 [7]	80.73	356.74	49.60	1.59	20.16	81.26	85.67	87.29	78.78	73.52	91.96
UNETR++ [25]	42.62	53.99	26.69	0.50	37.47	80.49	82.40	86.72	77.16	76.09	92.39
SegFormer3D [17]	4.50	**5.03**	**3.93**	0.17	**254.42**	74.34	81.66	86.57	73.85	55.07	91.64
SlimUNETR [15]	1.79	20.17	8.90	**0.12**	112.37	72.56	80.42	82.70	72.66	67.66	90.42
nnUNet [10]	31.78	417.96	23.97	1.29	41.72	77.82	82.11	84.57	74.37	53.14	91.89
3D UX-Net [12]	53.01	632.25	48.11	1.44	20.79	79.74	85.33	87.28	78.58	71.46	92.03
MedNeXt-S_K3 [24]	5.55	60.48	35.58	1.21	28.11	81.38	85.21	87.12	78.50	72.86	92.32
MedNeXt-L_K5 [24]	62.99	251.09	90.84	1.83	11.01	81.94	85.96	87.21	78.81	74.21	92.43
EfficientMedNeXt-T **(Ours)**	**0.43**	7.09	12.50	0.58	80.00	79.95	84.51	**87.86**	78.12	77.15	92.32
EfficientMedNeXt-S **(Ours)**	0.92	12.46	12.90	0.59	50.26	81.28	85.84	87.47	78.51	76.13	92.68
EfficientMedNeXt-M **(Ours)**	2.17	33.34	33.94	1.17	29.46	82.44	86.59	87.56	79.12	75.58	92.81
EfficientMedNeXt-L **(Ours)**	2.19	57.68	73.62	1.40	13.58	**83.29**	**87.00**	87.68	**79.21**	**77.31**	**92.98**

Table 3. Experimental results comparison of SOTA methods: *Params* (M), FLOPs (G), *Inference Time* (Inf. Time (ms)), GPU *Memory* (Mem. (GB)), and *Throughput* (Thrgh. (/s)) on BTCV13, BTCV8, FeTA, MSD Task01 Brain Tumour (BraT), Task06 Lung (Lung), and Task02 Heart (Heart) datasets. We retrained all 12 baseline models using their published architectures under the publicly available training framework and protocol from 3D UX-Net [12] (see their Appendix A.1, Table 4, GitHub) and EffiDec3D [21] (see their Supplementary Table S1). This setup ensures fair architecture-level comparisons without framework-specific advantages. We manually tuned all hyperparameters (e.g., learning rates) per each baseline's original guidelines for optimal results. We evaluated the 3D Generic_UNet (base=48) from nnUNet v1, without its adaptive pre/post-processing in order to maintain a uniform pipeline and compare only the architecture contribution. The *Inf. Time* and *Thrgh.* are reported for only the forward pass of a $96 \times 96 \times 96$ input averaging over 200 iterations on a NVIDIA RTX 6000 (Ada) GPU with 48GB memory. *Mem.* is the allocated peak GPU memory by Pytorch during the forward pass. The DICE scores (%) are reported averaging over five runs, thus having 0.5-3.5% standard deviations across datasets. Best values are in **bold**.

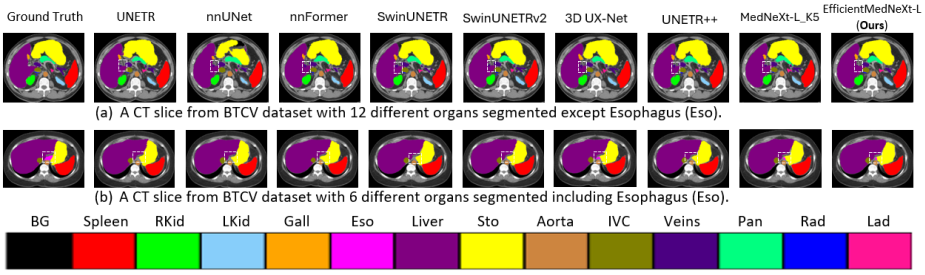

Fig. 2. Qualitative results of 13-organ segmentation on BTCV dataset. The dashed white box highlights incorrect predictions by most methods, including ours. **Note**: BG: background, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Sto: stomach, IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, Lad: left adrenal glands, Rad: right adrenal glands.

vein segmentation (white box) in all SOTA methods except our EfficientMedNeXt-L. Our model also achieves pixel-level precision for small structures (gallbladder, Lad, pancreas) and complex interfaces (IVC, veins), avoiding over-segmentation

artifacts seen in UNETR, nnFormer and 3D UX-Net. In Fig. 2b, only CNN-based architectures (3D UX-Net, MedNeXt-L_K5, our EfficientMedNeXt-L) segment the esophagus (white box), a critical yet frequently missed structure for radiotherapy planning. Our method uniquely balances holistic precision (e.g., stomach/aorta) with sub-millimeter boundary alignment (liver, kidneys), thus resolving persistent clinical workflow bottlenecks.

5 Conclusion

We have introduced EfficientMedNeXt, a computationally efficient segmentation architecture that balances segmentation performance and efficiency through Dilated Multi-Receptive Field Blocks (DMRFBs) and decoder optimization. Extensive experiments on BTCV, FeTA, and MSD BrainTumour, MSD Heart, and MSD Lung confirm its superiority over prior CNN- and Transformer-based models. With multiple network variants, EfficientMedNeXt offers scalable solutions for real-time and resource-constrained medical imaging.

Acknowledgments. This work is supported in part by the NSF grant CNS 2007284, in part by the iMAGiNE Consortium (https://imagine.utexas.edu/), and in part by the Texas Health Catalyst award.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European Conference on Computer Vision Workshops. pp. 205–218 (2022)
3. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., et al.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis p. 103280 (2024)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
6. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
7. He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D.: Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 416–426. Springer (2023)

8. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
9. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Transactions on Medical Imaging **42**(5), 1484–1494 (2023)
10. 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)
11. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. p. 12 (2015)
12. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. International Conference on Learning Representations (2023)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (2018)
15. Pang, Y., Liang, J., Huang, T., Chen, H., Li, Y., Li, D., Huang, L., Wang, Q.: Slim unetr: scale hybrid transformers to efficient 3d medical image segmentation under limited computational resources. IEEE Transactions on Medical Imaging **43**(3), 994–1005 (2023)
16. Payette, K., Li, H.B., de Dumast, P., Licandro, R., Ji, H., Siddiquee, M.M.R., Xu, D., Myronenko, A., Liu, H., Pei, Y., et al.: Fetal brain tissue annotation and segmentation challenge results. Medical image analysis **88**, 102833 (2023)
17. Perera, S., Navard, P., Yilmaz, A.: Segformer3d: an efficient transformer for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4981–4988 (2024)
18. Rahman, M.M., Marculescu, R.: Medical image segmentation via cascaded attention decoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6222–6231 (January 2023)
19. Rahman, M.M., Marculescu, R.: Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation. In: Medical Imaging with Deep Learning. pp. 1526–1544 (2023)
20. Rahman, M.M., Marculescu, R.: G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7728–7737 (2024)
21. Rahman, M.M., Marculescu, R.: Effidec3d: An optimized decoder for high-performance and efficient 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10435–10444 (2025)
22. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11769–11779 (2024)
23. 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)

24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
25. Shaker, A.M., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. IEEE Transactions on Medical Imaging (2024)
26. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 36–46. Springer (2021)
27. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119 (2021)
28. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
29. Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., et al.: Unest: local spatial representation learning with hierarchical transformer for efficient medical segmentation. Medical Image Analysis **90**, 102939 (2023)
30. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing **32**, 4036–4045 (2023)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3–11. Springer (2018)