HDNeXt: Hybrid Dynamic MedNeXt with Level Set Regularization for Medical Image Segmentation

Haoyu Cao, Tianyi Han, and Yunyun Yang*

School of Science, Harbin Institute of Technology, Shenzhen, China.
{22S058014, 22S058009}@stu.hit.edu.cn,
yangyunyun@hit.edu.cn

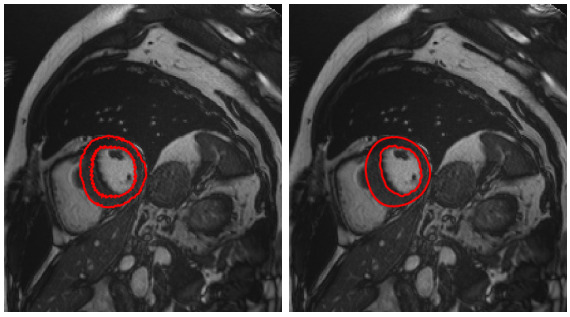
Abstract. Deep learning has been extensively employed in the field of medical image segmentation, demonstrating its robustness and efficacy. However, the pursuit of consistent segmentation performance across diverse instrumental conditions and the challenge of achieving precise boundary delineation in segmented images remain significant hurdles. In this paper, we aim to develop a model capable of achieving consistent, high-quality segmentation of identical regions of interest across varying instrumental conditions, with precise boundary delineation. Toward this end, we introduce our Hybrid Dynamic MedNeXt (HDNeXt) model, an advanced framework capable of dynamically generating weights across diverse medical images to maintain consistently high segmentation performance. HDNeXt builds on the robust segmentation framework of MedNeXt by incorporating dynamic convolution techniques, which endow the model with the capability for dynamic weight adjustment, significantly enhancing its segmentation performance. To tackle the second challenge, we devised a novel loss function, L_{CR}, formulated on the Curvature of the segmentation boundary and Region-Fitting energy derived from level set methods, which significantly enhances boundary precision during training and optimizes overall segmentation performance. Experiments were conducted on the abdominal CT datasets Synapse and the cardiac MRI datasets ACDC to demonstrate the efficiency and effectiveness of our method. Our method achieved an average Dice coefficient of 84.38 on the Synapse datasets and 93.59 on the ACDC datasets, surpassing other 2D state-of-the-art segmentation models and achieving optimal performance for 2D medical image segmentation. Codes are available at https://github.com/HaoyuCao/HDNeXt

Keywords: Medical Image Segmentation · Dynamic Networks · Loss Function Regularization.

1 Introduction

Medical image segmentation is a fundamental task in medical image processing. This task involves partitioning medical images into distinct regions that corre-

* Corresponding author.



(a) Serrated edges (b) Smooth edges

Fig. 1: Two images with similar Dice coefficients for segmentation are presented. Sub-figure (a) exhibits serrated edges that do not conform to anatomical structures. Sub-figure(b) features regular segmentation edges that meet clinical criteria, optimized through our L_{CR} loss function.

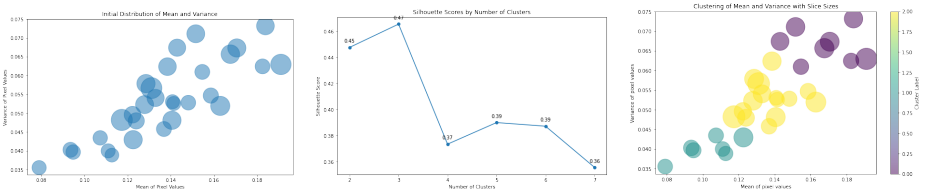
spond to various anatomical structures or regions of interest (ROIs). The advent of deep learning technologies has led to significant advancements in the accuracy and efficiency of medical image segmentation, establishing deep learning as the predominant approach in this field.

Generally, Convolutional Neural Networks (CNNs) are the fundamental deep learning approach for medical image segmentation. CNNs have demonstrated remarkable capabilities in image analysis tasks due to their hierarchical feature extraction process, exemplified by architectures such as the U-Net family[33,45,14] and the DeepLab series[5,35,22]. However, CNNs inherently suffer from a restricted receptive field due to window-based convolution calculations.

Recently, Vision Transformers (ViT)[8,24,23], initially developed for natural language processing, have been adapted to vision tasks with notable success. ViT is a model that directly applies the Transformer architecture[36] to sequences of image patches. It leverages the Transformer's capability to handle long-range dependencies and compensates for the inherent shortcomings of CNNs. Notable successes of Transformer models include TransUNet[4] and Swin-UNet[1] have demonstrated outstanding performance across various benchmarks.

A key reason for the notable success of Vision Transformers (ViT) in the visual domain is their strong scalability. However, in the field of medical imaging, where data annotation is costly, the development of ViT is severely constrained. Consequently, most modern networks predominantly use either purely CNN-based architectures or hybrid CNN-Transformer systems, capitalizing on the strengths of both.

Given these considerations, our study is based on the MedNeXt[34] architecture incorporating large kernel convolutions, a purely convolutional network design. Research has indicated that large kernel convolutions, can effectively integrate the strengths of convolutional networks and Transformers, thereby enhancing the network's feature extraction capabilities in image segmentation.



(a) Original distribution (b) Cluster scores (c) Cluster result

Fig. 2: The heterogeneity and cluster analysis of the Synapse dataset

Table 1: Different Synapse dataset partitions were obtained from the cluster analysis (Figure 2c). MedNeXt shows obvious inconsistent performance across different dataset partitions.

Different Division \ Synapse Score	Dice↑	Hd95↓
Green, Yellow (Train) - Purple (Test)	77.58	28.83
Green, Purple (Train) - Yellow (Test)	74.28	26.61
Yellow, Purple (Train) - Green (Test)	76.17	30.08
Random uniform division	79.92	25.19

MedNeXt architecture has demonstrated superior performance across various medical image segmentation tasks, achieving state-of-the-art (SOTA) results on board and surpassing networks based on the three aforementioned architectures.

However, despite MedNeXt's outstanding performance across various metrics, our findings indicate that MedNeXt struggles to maintain consistent segmentation performance across datasets with varying distributions. As shown in Figure 2, we calculated the clustering scores of the Synapse dataset (Figure 2b) and discovered that the Synapse dataset can be divided into three clusters (Figure 2c). By using two clusters as the training set and one cluster as the testing set, as well as random uniform divisions, the performance of MedNeXt exhibits significant fluctuations (Table 1). Additionally, MedNeXt exhibits coarse and blurred boundary delineations due to a lack of precise boundary awareness during segmentation (Figure 1a).

In this work, we propose HDNeXt (Hybrid Dynamic MedNeXt), a segmentation model that achieves consistent segmentation performance while maintaining precise boundary delineations.

The major contributions of our work can be summarized as follows:

1. We developed the novel Dynamic MedNeXt Module (DyNeXt), which incorporates dynamic convolutions to endows the network with dynamic weighting capabilities. DyNeXt augments both the efficacy and uniformity of network-based segmentation outcomes.

2. We designed a loss function L_{CR} based on curvature regularization and region fitting term in level set frame work, which significantly enhances the accuracy of boundary segmentation and smooths the segmentation results.
3. Based on these innovations, we proposed the novel HDNeXt network architecture and conducted experiments on Synapse and ACDC datasets. Our network achieved state-of-the-art performance in 2D medical image segmentation.

2 Related Work

2.1 Large Kernel Convolutional Neural Networks

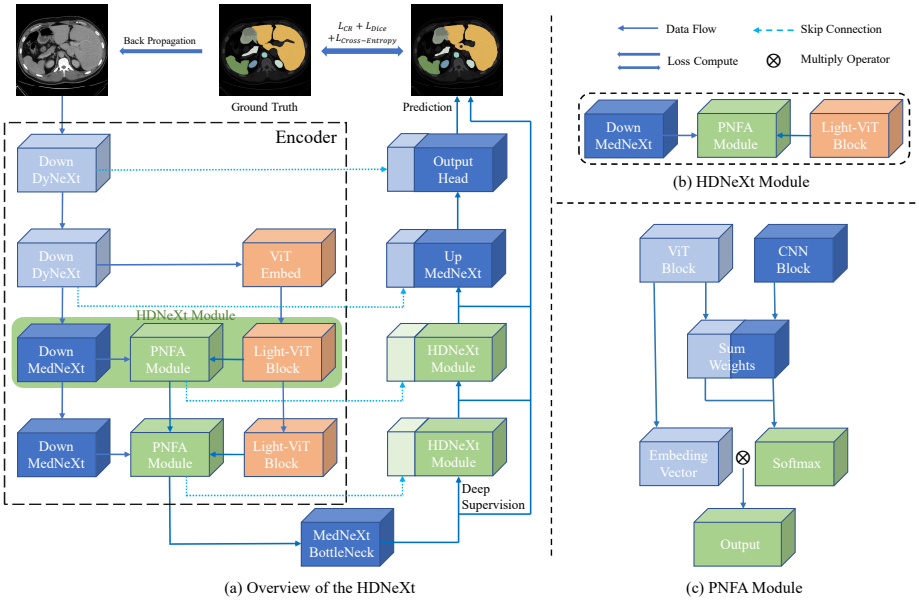
The development of large kernel convolutional neural networks has played a key role in overcoming the limitations of traditional CNN architectures, particularly in enhancing the receptive field and more effectively capturing global context. The seminal work by Peng et al.[28] highlights the utility of large kernels in bridging the gap between pixel-level predictions and global contextual understanding necessary for accurate semantic segmentation.

The ConvNeXt series[25,37,43] employs larger convolutional kernels, such as 7x7 or even larger, to increase the receptive field. These models incorporates design ideas from the architecture of Transformers, adjusting the scale of layers, regularization strategies, pre-training methods, and kernel sizes, thereby demonstrating that pure convolutional networks can achieve performance on pair with or even surpassing that of Transformers in visual tasks. MedNeXt[34] adapts this large kernel convolutional network to medical image segmentation, showcasing its immense potential in handling complex 3D medical image segmentation tasks and establishing a new state-of-the-art model.

2.2 Dynamic Weight Network

Making the weights of a neural network sample-adaptive through dynamic mechanisms has shown great potential for boosting model capacity and generalisation. DyConv[6] enables dynamic adjustment of weights by generating specific convolution kernels for each input. CondConv[41] increases the capacity and flexibility of the model by using different convolution kernels for different input conditions. OmniConv[19] proposes an all-inclusive convolution strategy that dynamically adjusts the size and shape of the convolution kernel according to the characteristics of the input.

In our work, we discovered that the MedNeXt model does not achieve consistent segmentation performance across images from different imaging conditions. To address this issue, we leverage the advantages of dynamic weight networks and propose a large kernel convolution module based on dynamic convolutions. This module enables the network to adaptively generate dynamic weights based on the input, thereby enhancing its generalization and segmentation capabilities.



(a) Overview of the HDNeXt (b) HDNeXt Module (c) PNFA Module

Fig. 3: Overview of our HDNeXt model. The proposed HDNeXt is a nested double-U model with a large size CNN branch and an auxiliary light weight ViT branch. The PNFA module plays a significant role in the feature fusion of CNN-ViT without adding extra learnable parameters.

2.3 Level Set Regularization

The level set method, a mathematical algorithm for segmentation via iterative evolution of high-dimensional functions, is often used as a regularization term in deep learning due to its adaptive handling of topological changes. Kim et al.[16] were the first to embed the level set function as a loss term in neural network training and demonstrated the great potential of the level set method as a regularization term. Kim and Ye [15] extended the Mumford–Shah functional as a loss function effectively captures smooth segmentation boundaries. Yang et al.[42] were the first to embed the curvature-based euler functional into the network as a loss function, significantly enhancing network performance. Researchers derived different loss functions by designing energy functionals with varying properties, further regularizing and enhancing model performance. Based on the Euler elasticity model, we proposed a curvature and region based loss L_{CR}, which effectively optimizes boundary accuracy and smoothness.

3 Methodology

As shown in Figure 3, the proposed HDNeXt model is a nested double-U structure comprising a large-size CNN branch and a small-size light weight Trans-

former branch in Figure 3. The network uses CNN as the fundamental backbone, serving as the source for processing inputs and generating segmentation masks. The original image is downsampled through two layers of DyNeXt modules. The DyNeXt modules not only enable the network to adaptively adjust weights based on the input but also generate more features in the shallow layers of the network.

Then we divide the downsampled image into patches and perform token embedding to introduce a light weight auxiliary ViT branch. The light weight ViT component does not directly participate in the network's predictions but serves as an auxiliary branch to establish long-range dependencies, compensating for the limited receptive field of the convolutional network. The CNN-ViT interacts through our innovatively designed PNFA module in Figure 3, which enables effective fusion of dual-path features without adding extra learnable parameters to the network.

The hybrid loss incorporating L_{CR} is computed on the final predicted segmentation masks and updates the model parameters through back-propagation. We will describe the details of our model design in the following sections and demonstrate the effectiveness of our module design through comparative and ablation experiments.

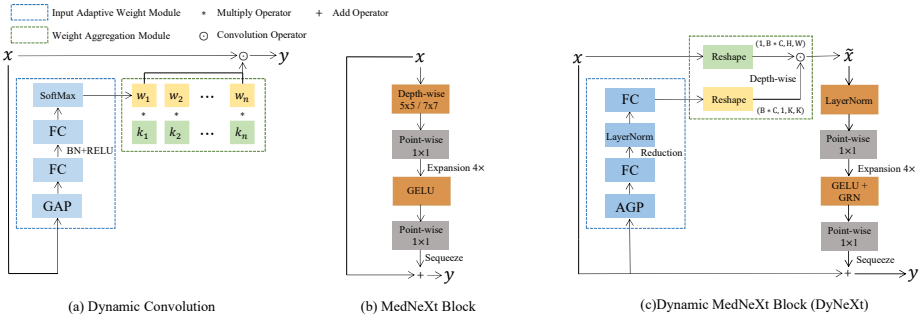
3.1 Dynamic MedNeXt Module(DyNeXt)

Dynamic weight networks have been proven to possess good generalization capabilities when dealing with non-uniformly distributed datasets. we propose our Dynamic MedNeXt(DyNeXt) Module as shown in Figure 4 building on the concept of dynamic modules[6,32,9] to introduce dynamic properties.

Our implementation focuses on the MedNeXt[34] module (see Figure 4). MedNeXt module follows the design philosophy of the large kernel convolution, comprising a depth-wise convolution using a 7×7 kernel, as well as two 1×1 convolutions in an inverted bottleneck design. We assume the input is in the shape of (B, C, H, W). Dynamic properties can be divided into two modules:

The first module is a self-operating branch, referred to as the Input Adaptive Weight Module. It generates parameters for depth-wise convolution through adaptive global pooling followed by two fully connected layers based on the input. The second module is called the Weight Aggregation Module, where the reshaped original input undergoes self-convolution operations with the weights generated by the first module.

The computation flow of DyNeXt can be expressed with the following equations. Let $\tilde{x}$ be the output of the Input Adaptive Weight Module, and x' be the output after the Weight Aggregation Module. The two outputs are combined with the residual from the original input through an invert bottleneck, resulting



(a) Dynamic Convolution (b) MedNeXt Block (c)Dynamic MedNeXt Block (DyNeXt)

Fig. 4: Overview of DyNeXt Block. We introduce dynamic properties into the 2D MedNeXt Block, endowing the network with dynamic weight properties. We add a self-operating module called the Input Adaptive Weight Module in MedNeXt, which dynamically generates depth-wise convolution kernel weights based on the input and performs self-convolution operations with the original input.

in the final output of the DyNeXt module. The equations are as follows:

$$\tilde{x} = \mathrm{FC}(\mathrm{LN}(\mathrm{FC}(\mathrm{AGP}(x)))) \tag{1}$$

$$x' = \underbrace{\mathrm{reshape}(x)}_{(1,B\times C,H,W)} \odot \underbrace{\mathrm{reshape}(\tilde{x})}_{(B\times C,1,K,K)} \tag{2}$$

$$y = (\mathrm{InvertBottleneck}(\mathrm{LN}(x'))) + x \tag{3}$$

Through the interaction of the two modules, the DyNeXt module adaptively generates convolution weights based on the input, producing input-adaptive outputs.

3.2 Light-ViT Block

To address the limited receptive field in CNNs, we introduce an auxiliary Lightweight ViT branch at an appropriate downsampling stage. This branch helps establish long-range dependencies, enhancing the model's ability to segment abdominal CT scans, particularly when target objects span multiple regions or have elongated structures. This approach significantly improves the model's capacity to capture spatial relationships crucial for accurate segmentation.

The Light-ViT structure is illustrated in Figure 5. We employ the original Swin Transformer [24] to reduce the self-attention computational complexity from N^2 to linear. In medical image segmentation, the frequency domain often captures higher-dimensional semantic information with reduced noise, which is critical for clinical applications where detecting signal amplitude variations helps differentiate tissues and organs. Thus, we incorporate the principles of LightViT [11] with spatial frequency domain extraction [40,13,38] in the feed-forward neural network. Specifically, feature extraction is conducted in the spatial frequency domain using Fast Fourier Transform (FFT) and its inverse operation (iFFT)

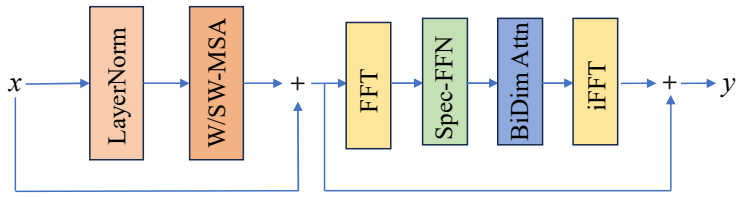


Fig. 5: Overview of our Light-ViT Block. We adopt the Swin Transformer as the foundation module, applying the feed-forward neural network component in the frequency space to capture high-level semantic information.

for reconstruction. This integration optimizes processing speed and scalability, making the transformer architecture more suitable for efficient segmentation with enhanced contextual understanding.

3.3 Pre-Normalized Feature Aggregation(PNFA) Module

We further explored the distinctions between Transformer modules and dynamic convolutional networks. The computation in a deep convolutional network can be expressed as follows:

$$y_i = \sum_{j \in \mathcal{L}(i)} w_{i-j} \odot x_j, \tag{4}$$

Self-attention is the fundamental computational unit of the Transformer architecture. Its computation can be represented by the following formula:

$$y_i = \sum_{j \in \mathcal{L}(i)} \frac{\exp\left(x_i^\top x_j\right)}{\sum_{k \in \mathcal{L}(i)} \exp\left(x_i^\top x_k\right)} x_j \tag{5}$$

An ideal network model should combine the inherent translation invariance of convolutional neural networks with the dynamic weighting and global receptive field capabilities of Transformer networks. Our method is inspired by CoAtNet[7,10] for feature aggregation involves directly summing a global depth-wise convolutional kernel with an adaptive attention matrix before the Softmax normalization. This approach results our method of pre-normalized feature aggregation (PNFA):

$$y_i^{\text{pre}} = \sum_{j \in \mathcal{L}(i)} \frac{\exp\left(x_i^\top x_j + w_{i-j}\right)}{\sum_{k \in \mathcal{L}(i)} \exp\left(x_i^\top x_k + w_{i-k}\right)} x_j, \text{(Pre-Normalization)} \tag{6}$$

We discovered that the pre-normalized feature aggregation module corresponds to a specific variant of the relative self-attention mechanism[36]. In this scenario, the attention weights are jointly determined by convolutional inputs, which exhibit translational invariance, and by adaptively weighted inputs. The output is capable of simultaneously leveraging the complementary features of both aspects. The PNFA module is illustrated in Figure 3.

3.4 Curvature-Region Regularization Loss

The level set method possesses excellent properties for handling topological changes, which can effectively compensate for the insensitivity of neural networks to boundaries when used as a loss function. We consider a class of Euler elasticity models from the level set energy functional:

$$\min \underbrace{\int_{\Omega} \psi(\kappa)|\nabla u| \mathrm{d}x}_{\mathcal{R}(\cdot)} + \mu \mathcal{D}(u, f), \tag{7}$$

where $\mathcal{R}(\cdot)$ is the curvature regularity term, and $\mathcal{D}(u, f)$ the data fidelity term. In the level set method, $\mathcal{R}(\cdot)$ measures the smoothness of segmentation edges, while $\mathcal{D}(u, f)$ ensures convergence to the ground truth. Following Eqn. (7), we use the ADMM and augmented Lagrange Multiplier methods to minimize the energy functional, as shown in Figure 6:

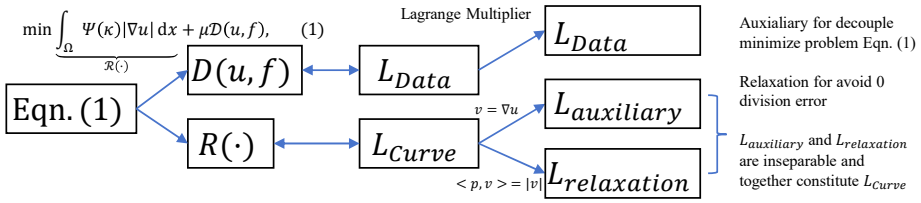


Fig. 6: Derivation Process of L_{CR} via ADMM and augmented Lagrange method.

We choose TSC [18] for $\mathcal{R}(\cdot)$ and RSF [21] for $\mathcal{D}(u, f)$. The rationale for these selections is explained in Section 4.4. During the optimization process, several new parameters are introduced, u represents the input image, and ∇ denotes the gradient operator. We introduce a relaxation variable $v = \nabla u$ to decouple the problem, and a auxiliary variable p, such that $< p, v >= |v|$ to prevent division by zero errors. Δ refers to the Laplace operator, while w serves as the multiplier in the augmented Lagrange method. μ, β, α, η and λ_i are non-negative constant parameters. ϕ is the level set function, and $\delta(\cdot)$ and e_i follows the definition in the original RSF[21] paper. Finally, we embed the equivalent first-order optimality condition of Eqn. (7) in residual form as a loss function into the neural network:

$$\begin{cases} L_{\text{Data}} &= \eta \Delta u - \nabla \cdot [w + v] + \mu \delta(\phi) \left[-\lambda_1 e_1 + \lambda_2 e_2\right] \\ L_{\text{Auxiliary}} &= \nabla(-\eta u + 2\alpha \nabla \cdot u) + w + \eta v \\ L_{\text{Relaxation}} &= 2\beta \nabla(\nabla \cdot p)|v| - \nabla u \end{cases} \tag{8}$$

Our final curvature-region regularization loss is composed of the above three residual forms of the first order optimality conditions:

$$L_{CR} = L_{\text{Data}} + \underbrace{L_{\text{Auxiliary}} + L_{\text{Relaxation}}}_{L_{Curve}}$$

Table 2: Performance comparison between our HDNeXt and other state-of-the-art methods on the ACDC dataset

Method	Backbone	Resolution	Dice↑	IoU↑	Hd95↓
UNet++	CNN	512^2	89.01	80.43	9.24
nnUNet	CNN	512^3	92.61	86.52	8.63
MedNeXt	CNN	128^3	91.36	84.41	5.91
TransUNet	Hybrid	224^2	90.71	84.26	8.77
PVT-CASCADE	Hybrid	224^2	89.62	83.03	7.32
TransCASCADE	Hybrid	224^2	89.07	82.84	8.99
PVT-GCASCADE	Hybrid	224^2	92.46	87.12	2.58
Swin-UNet	Transformer	224^2	91.00	84.39	3.31
MERIT	Transformer	256^2	91.58	86.21	4.81
MISSFormer	Transformer	224^2	88.90	83.66	5.34
nnFormer	Transformer	512^3	92.78	87.04	2.37
HDNeXt	Hybrid	224^2	**93.59**	**87.62**	**1.53**

Initially, we do not introduce the L_{CR} loss function. Once the network produces stable predictions, L_{CR} is incorporated to further regularize and optimize the network. At this stage, the holistic loss function of the model consists of Dice Loss, Cross-Entropy Loss, and L_{CR}, with equal weights assigned to each term:

$$L_{Total} = L_{Dice} + L_{Cross-Entropy} + L_{CR}$$

4 Experiment

We evaluated HDNeXt's performance on image segmentation using the Dice coefficient, IoU score for overall segmentation accuracy, and the 95% Hausdorff distance (Hd95) to assess boundary precision. We use abdominal CT datasets Chaos and cardiac MRI datasets Cardiac as auxiliary datasets for SimMIM pretraining strategy [39].

Comprehensive comparisons across these datasets were conducted against state-of-the-art models, including CNN-based architectures (UNet++ [45], nnUNet [14], MedNeXt (2D) [34]), Transformer models (Swin-UNet [1], MERIT [31], MISSFormer [12], nnFormer [44]), and hybrid architectures (TransUNet [4], PVT-CASCADE [29], TransCASCADE [29], PVT-GCASCADE [30]).

4.1 ACDC Dataset Comparison Experiment

Table 2 compares the performance of HDNeXt with various state-of-the-art segmentation models on the ACDC dataset, showcasing HDNeXt's superior performance.mong the evaluated models, HDNeXt, a hybrid architecture, demonstrates the highest Dice score of 93.59% and the highest IoU of 87.62%, beating

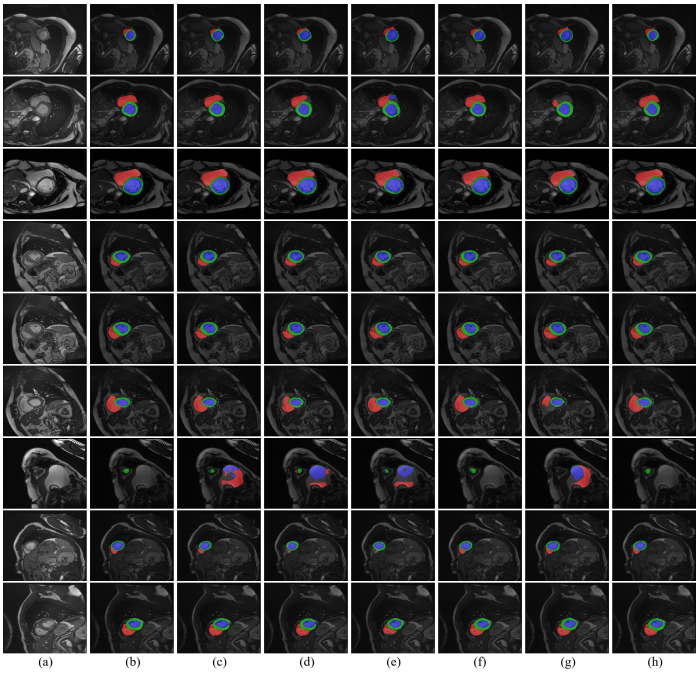


Fig. 7: Qualitative results on ACDC dataset. (a) Origin image. (b) Ground truth. (c) MedNeXt. (d) PVT-CASCADE. (e) PVT-GCASCADE. (f) MISSFormer. (g) nnFormer. (h) Ours.

nnFormer by 0.81% and 0.58% respectively. HDNeXt also achieves the lowest Hd95 index of 1.53mm^2, which indicates more precise contour delineations compared to its counterparts. For instance, PVT-GCASCADE and the transformer-based nnFormer model record higher Hausdorff distances of 2.58 mm^2 and 2.37 mm^2, respectively, suggesting less accuracy in capturing boundary details.

We selected competitive models for visualization based on their performance metrics on the ACDC dataset. As shown in Figure 2, visible results illustrates the comparison of our segmentation results with those of other models under different conditions (ES/ED) and various instrument settings. The visualization results clearly demonstrate that our segmentation achieves superior overall and edge accuracy, as well as enhanced stability.

4.2 Synapse Dataset Comparison Experiment

Table 3 compares HDNeXt with leading segmentation models on the Synapse dataset. HDNeXt achieves a Dice coefficient of 84.38%, slightly surpassing MERIT (84.03%) and indicating improved segmentation consistency. Notably, HDNeXt also excels in boundary accuracy, with a significantly lower Hd95 value of 10.53mm^2 compared to PVT-GCASCADE's 13.23mm^2. PVT-CASCADE and TransCAS-

Fig. 8: Qualitative results on Synapse dataset. (a) Origin image. (b) Ground truth. (c) MedNeXt. (d) PVT-CASCADE. (e) PVT-GCASCADE. (f) MISSFormer. (g) nnFormer. (h) Ours.

CADE, achieve Dice scores of 81.06% and 82.68% but fall short in boundary accuracy, with Hd95 values of 20.23 mm^2 and 17.34 mm^2.

Figure 8 visualizes segmentation results on the Synapse dataset, sampling three longitudinal abdominal CT slices (high, medium, low positions). Thanks to the DyNeXt module, HDNeXt consistently outperforms other models, especially on challenging Synapse samples.

4.3 Ablation Study on Network Components

Table 4 illustrates the ablation study results for the HDNeXt model variants on both Synapse dataset and ACDC dataset. The baseline model MedNeXt (2D Version) achieves a Dice score of 79.92% and a Hd95 of 25.19 mm^2 on Synapse and a Dice score of 91.08% and a Hd95 of 6.33 mm^2. This serves as the foundational performance metric for further enhancements.

The integration of the DyNeXt module yields a substantial performance boost, raising the Dice score to 82.98% and reducing Hd95 to 19.49 mm^2 on

Table 3: Performance comparison between our HDNeXt and other state-of-the-art methods on the Synapse dataset

Method	Backbone	Dice↑	Hd95↓
UNet++	CNN	76.03	33.26
nnUNet	CNN	76.63	25.26
MedNeXt	CNN	80.62	22.68
TransUNet	Hybrid	77.48	31.69
PVT-CASCADE	Hybrid	81.06	20.23
TransCASCADE	Hybrid	82.68	17.34
PVT-GCASCADE	Hybrid	83.06	13.23
Swin-UNet	Transformer	79.13	21.55
MERIT	Transformer	84.03	14.52
MISSFormer	Transformer	81.96	18.20
HDNeXt	Hybrid	**84.38**	**10.53**

Table 4: Network Components Ablation studies on the Synapse and ACDC.

								Synapse		ACDC	
Method	SimMIM	CNN	DyNeXt	ViT	PNFA	Skip	L_{CR}	Dice↑	Hd95↓	Dice↑	Hd95↓
MedNeXt		✓				✓		79.92	25.19	91.08	6.33
	✓	✓				✓		80.53	22.68	91.36	5.91
HDNeXt		✓	✓			✓		82.98	19.49	92.51	5.34
	✓	✓	✓	✓		✓		83.37	22.31	92.63	5.08
	✓	✓	✓	✓	✓	✓		83.82	15.52	93.03	4.67
	✓	✓	✓	✓	✓			83.48	19.36	92.59	4.98
	✓	✓		✓	✓	✓		83.53	22.47	92.67	5.14
HDNeXt $+L_{CR}$	✓	✓	✓	✓	✓	✓	✓	**84.38**	**10.53**	**93.59**	**1.53**

Synapse, while enhancing the Dice score to 92.51% and lowering Hd95 to 5.34 mm^2 on ACDC. The addition of the Vision Transformer (ViT) as an auxiliary feature extraction branch, coupled with the PNFA module for feature fusion, further amplifies model performance. The complete HDNeXt configuration, including SimMIM, DyNeXt, ViT, PNFA, and skip connections, achieves a Dice score of 83.82% and a Hd95 of 15.52 mm^2 on Synapse, along with a Dice score of 93.03% and a Hd95 of 4.67 mm^2, underscoring its robust feature representation and effective feature integration capabilities.

The HDNeXt + L_{CR} configuration further refines the Dice coefficient to 84.38% with a notable reduction in Hd95 to 10.53 mm^2 on the Synapse dataset, along with improvements in the Dice score to 93.59% and Hd95 to 1.53 mm^2 on the ACDC dataset. underscoring L_{CR}'s efficacy in enhancing boundary delineation accuracy and achieving significant advancements in edge refinement.

Table 5: Comparison Study on Curvature Regularization Term TAC, TRV, and TSC (Left) and Data Fidelity Term: CV, RSF, MICO, ALF, and RLSF (Right).

Model	Curvature Loss	Synapse Dice↑	Synapse Hd95↓	ACDC Dice↑	ACDC Hd95↓	Model	Data Term	Synapse Dice↑	Synapse Hd95↓	ACDC Dice↑	ACDC Hd95↓
HDNeXt	—	83.82	15.52	93.03	4.67	HDNeXt $+L_{Curve}$	—	84.04	12.77	93.19	2.37
	+TAC	83.99	13.91	93.18	2.84		+CV	83.25	14.97	91.80	4.18
	+TRV	84.02	16.32	**93.21**	3.63		**+RSF**	84.38	**10.53**	**93.59**	1.53
	+TSC	**84.04**	**12.27**	93.19	**2.37**		+MICO	**84.46**	14.81	93.33	1.25
							+ALF	84.45	11.37	93.39	**1.07**
							+RLSF	83.90	22.03	93.53	2.14

4.4 Ablation Study on L_{CR} Components

We conducted an ablation study on the components of L_{CR}, exploring different curvature regularization strategies (TAC[17], TRV[2], TSC[18]) and data fidelity terms (C-V[3], RSF[21], MICO[20], ALF[26], and RLSF[27]), with results summarized in Table 5. The results indicate that selecting TSC as the curvature term and RSF as the data fidelity term yielded the best performance on both the Synapse and ACDC datasets, resulting in an increase in Dice score from 83.82% to 84.38% and a reduction in Hd95 from 15.52 mm^2 to 10.53 mm^2 on the Synapse dataset and a Dice score improved from 93.03% to 93.59%, and the Hd95 index decreased from 4.67 mm^2 to 1.53 mm^2 on the ACDC dataset.

It is important to note that due to the flexibility of the variational-based energy functional in Eqn. (7), these choices are not fixed. Different combinations may outperform the current configuration. Our selection was primarily based on experimental outcomes. Further research in this area holds significant value.

5 Conclusion

In this paper, we present HDNeXt, a segmentation network featuring the DyNeXt module, a dynamically adaptive large-kernel convolution that enhances segmentation performance and consistency. We also propose a meticulously designed loss function, L_{CR}, based on curvature regularization and regional energy terms, which improves boundary accuracy and can be viewed as a plug-in loss function applicable to other segmentation tasks.

Our extensive experiments substantiate that HDNeXt attains state-of-the-art performance in 2D medical image segmentation, highlighting the efficacy of our proposed methodologies and contributions. We intend to further evaluate the method's robustness across diverse datasets and aim to extend L_{CR} to a 3D formulation, facilitating its application to 3D segmentation in future work.

Acknowledgments. This research is supported by National Natural Science Foundation of China No. 62371156 and Natural Science Foundation of Guangdong Province No. 2022A1515011629, University Innovative Team Project of Guangdong 2022KCXTD039, and National Natural Science Foundation of China (12371419).

References

1. 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: European Conference on Computer Vision. pp. 205–218. Springer (2022)
2. Chambolle, A., Pock, T.: Total roto-translational variation. Numerische Mathematik **142**, 611–666 (2019)
3. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing **10**(2), 266–277 (2001)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. ArXiv preprint arXiv:2102.04306 (2021)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(4), 834–848 (2017)
6. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11030–11039 (2020)
7. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. Advances in Neural Information Processing Systems **34**, 3965–3977 (2021)
8. 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)
9. Ha, D.T., Phuong, D.L.: Freedom of information law comes to vietnam: How do human rights adapt to goals of economic development and political stability? Austl. J. Asian L. **18**, 167 (2017)
10. Han, Q., Fan, Z., Dai, Q., Sun, L., Cheng, M.M., Liu, J., Wang, J.: On the connection between local attention and dynamic depth-wise convolution. ArXiv preprint arXiv:2106.04263 (2021)
11. Huang, T., Huang, L., You, S., Wang, F., Qian, C., Xu, C.: Lightvit: Towards light-weight convolution-free vision transformers. arXiv preprint arXiv:2207.05557 (2022)
12. Huang, X., Deng, Z., Li, D., Yuan, X.: Missformer: An effective medical image segmentation transformer. arXiv preprint arXiv:2109.07162 (2021)
13. Huang, Z., Zhang, Z., Lan, C., Zha, Z.J., Lu, Y., Guo, B.: Adaptive frequency filters as efficient global token mixers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6049–6059 (2023)
14. 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)
15. Kim, B., Ye, J.C.: Mumford–shah loss functional for image segmentation with deep learning. IEEE Transactions on Image Processing **29**, 1856–1866 (2019)
16. Kim, Y., Kim, S., Kim, T., Kim, C.: Cnn-based semantic segmentation using level set loss. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1752–1760. IEEE (2019)
17. Kuiper, N.H.: Minimal total absolute curvature for immersions. Inventiones mathematicae **10**(3), 209–238 (1970)

18. Langer, J., Singer, D.A.: The total squared curvature of closed curves. Journal of Differential Geometry **20**(1), 1–22 (1984)
19. Li, C., Zhou, A., Yao, A.: Omni-dimensional dynamic convolution. ArXiv preprint arXiv:2209.07947 (2022)
20. Li, C., Gore, J.C., Davatzikos, C.: Multiplicative intrinsic component optimization (mico) for mri bias field estimation and tissue segmentation. Magnetic Resonance Imaging **32**(7), 913–923 (2014)
21. Li, C., Kao, C.Y., Gore, J.C., Ding, Z.: Minimization of region-scalable fitting energy for image segmentation. IEEE Transactions on Image Processing **17**(10), 1940–1949 (2008)
22. Liu, C., Chen, L.C., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 82–92 (2019)
23. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019 (2022)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
25. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
26. Ma, D., Liao, Q., Chen, Z., Liao, R., Ma, H.: Adaptive local-fitting-based active contour model for medical image segmentation. Signal Processing: Image Communication **76**, 201–213 (2019)
27. Niu, S., Chen, Q., De Sisternes, L., Ji, Z., Zhou, Z., Rubin, D.L.: Robust noise region-based active contour model via local similarity factor for image segmentation. Pattern Recognition **61**, 104–119 (2017)
28. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4353–4361 (2017)
29. 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 (2023)
30. 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)
31. 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. PMLR (2024)
32. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. ArXiv preprint arXiv:1803.00676 (2018)
33. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)

34. 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)
35. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 843–852 (2017)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
37. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
38. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6030–6038 (2024)
39. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
40. Xu, K., Qin, M., Sun, F., Wang, Y., Chen, Y.K., Ren, F.: Learning in the frequency domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1740–1749 (2020)
41. Yang, B., Bender, G., Le, Q.V., Ngiam, J.: Condconv: Conditionally parameterized convolutions for efficient inference. Advances in Neural Information Processing Systems **32** (2019)
42. Yang, Y., Yan, T., Jiang, X., Xie, R., Li, C., Zhou, T.: Mh-net: Model-data-driven hybrid-fusion network for medical image segmentation. Knowledge-Based Systems **248**, 108795 (2022)
43. Yu, W., Zhou, P., Yan, S., Wang, X.: Inceptionnext: When inception meets convnext. arXiv preprint arXiv:2303.16900 (2023)
44. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. ArXiv preprint arXiv:2109.03201 (2021)
45. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019)