# Med-DANet: Dynamic Architecture Network for Efficient Medical Volumetric Segmentation

Wenxuan Wang[1], Chen Chen[2], Jing Wang[1], Sen Zha[1], Yan Zhang[1], and Jiangyun Li[1,†]

[1] School of Automation and Electrical Engineering, University of Science and Technology Beijing, China, s20200579@xs.ustb.edu.cn, m202120718@xs.ustb.edu.cn, g20198675@xs.ustb.edu.cn, m202110578@xs.ustb.edu.cn, leejy@ustb.edu.cn

[2] Center for Research in Computer Vision, University of Central Florida, USA, chen.chen@crcv.ucf.edu

[3] † Corresponding author: Jiangyun Li

**Abstract.** For 3D medical image (e.g. CT and MRI) segmentation, the difficulty of segmenting each slice in a clinical case varies greatly. Previous research on volumetric medical image segmentation in a slice-by-slice manner conventionally use the identical 2D deep neural network to segment all the slices of the same case, ignoring the data heterogeneity among image slices. In this paper, we focus on multi-modal 3D MRI brain tumor segmentation and propose a dynamic architecture network named Med-DANet based on adaptive model selection to achieve effective accuracy and efficiency trade-off. For each slice of the input 3D MRI volume, our proposed method learns a *slice-specific decision* by the Decision Network to dynamically select a suitable model from the predefined Model Bank for the subsequent 2D segmentation task. Extensive experimental results on both BraTS 2019 and 2020 datasets show that our proposed method achieves comparable or better results than previous state-of-the-art methods for 3D MRI brain tumor segmentation with much less model complexity. Compared with the state-of-the-art 3D method TransBTS, the proposed framework improves the model efficiency by up to 3.5× without sacrificing the accuracy. Our code will be publicly available at https://github.com/Wenxuan-1119/Med-DANet.

**Keywords:** Segmentation · Brain Tumor · MRI · Dynamic Network · Adaptive Inference

## 1 Introduction

Gliomas are the most common malignant brain tumors with different levels of aggressiveness. The precise measurements of gliomas can assist doctors in making accurate diagnosis and further treatment planning. Traditionally, the lesion regions are delineated by clinicians heavily relying on clinical experiences, which is time-consuming and prone to mistakes. Therefore, to improve the accuracy and

efficiency of clinical diagnosis, automated and accurate segmentation of these malignancies on Magnetic Resonance Imaging (MRI) [12] is of vital importance.

In the past few years, deep neural networks, convolutional neural networks (CNNs) in particular, have achieved great success in medical image segmentation task. The mainstream methods can be divided into two categories: (1) applying 2D networks for slice-wise (i.e. slice-by-slice) predictions and (2) utilizing 3D models (e.g. 3D CNNs) to process image volumes with multiple slices. 3D CNNs such as 3D U-Net [8] and V-Net [22] employing 3D convolutions to capture the correlation between adjacent slices, have achieved impressive segmentation results. However, these 3D CNN architectures come with high computational overheads due to multiple layers of 3D convolutions, making them prohibitive for practical large-scale applications. Similarly, the 2D U-Net [27] and its variants such as [24, 36, 37] are also confronted with the same problem because of the unique architecture. Specifically, to obtain the multi-scale feature representation and fine-grained local details, multiple skip connections and stacked stacked convolutional layers are employed to improve model performance, but leading to unbearable computational overheads simultaneously.

Since the efficiency of a network determines the practical application value of the model deployment, model efficiency is as important as segmentation accuracy. In order to cope with the high computational costs brought by 3D medical image itself and the segmentation networks mentioned above, many lightweight networks [4,7,16,18,23,26] have been developed to realize efficient medical image segmentation. However, these proposed lightweight networks are designed from the perspective of efficient architecture without the consideration of data itself, treating all different inputs equally. Although these models effectively make the structural improvements to achieve lightweight architectures, they suffer from segmentation accuracy degradation due to reduced modeling capacity. Moreover, they can not **adaptively** make appropriate adjustments to different input data due to fixed network structure. Therefore, a natural question arises:

*For volumetric medical image segmentation task, is it possible to achieve dynamic inference with adjustable network structures for better accuracy and efficiency trade-offs by considering the characteristics of the input data (e.g. the level of segmentation difficulty of each image slice)?*

To answer this question, we take a brain tumor segmentation dataset BraTS 2019 [1,2,21] as an example to seek some insights. Fig. 1 (a) shows the distribution of a 3D multi-modal brain tumor image along the slice dimension for one case. Due to several factors, such as the MRI process, shape of the organ (e.g. brain), and the location of the disease (e.g. glioma), the image content varies significantly across different MRI slices. For example, the $1^{st}$ row of Fig. 1 (a) shows the first 5 slices that barely capture any tissue content of the brain. These slices can be simply predicted as containing all "background" pixels (i.e. no lesion pixel) without model inference (i.e. "skip" mode), saving the computational cost. For MRI slices contain lesions, the level of difficulty for segmentation also varies a lot. Some slices contain only certain categories of the foreground or tumor morphology is easy to segment (as highlighted with blue boxes in Fig. 1

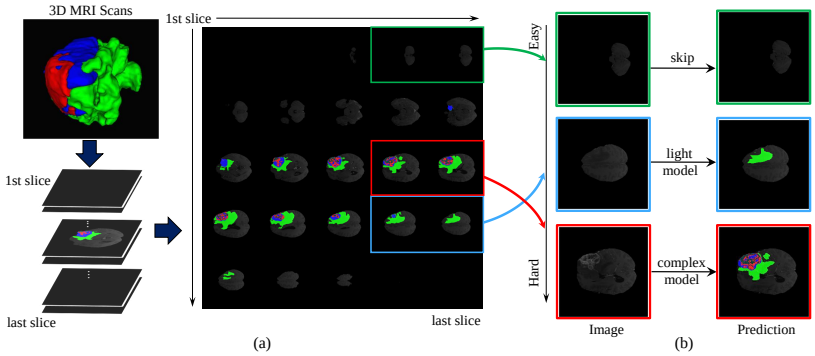



**Fig. 1.** (a) The illustration of image content distribution along slice dimension of an MRI case (Axial View) from the BraTS 2019 dataset. The blue regions denote the enhancing tumors, the red regions denote the non-enhancing tumors, and the green ones denote the peritumoral edema. (b) The main idea of our proposed framework for dynamic inference. For image slices with diverse segmentation difficulty, our framework realizes efficient and accurate segmentation by adaptively adjusting the architecture, selecting the optimal network in the Model Bank which consists of several networks with different model complexities. In this way, our framework can dynamically decide to "slack off" or "work hard" according to different samples.

(b)), and some difficult-to-segment slices contain multiple types of tumors that are extremely irregular in shape and difficult to recognize (as highlighted with red boxes in Fig. 1 (b)). From the analysis of the MRI data, the answer to the above question is Yes. It is possible to adjust the model complexity according to the input (e.g. image slice) for effective accuracy and efficiency trade-offs.

**Our Solution.** In this paper we tackle the aforementioned high computational overload problem of medical volumetric segmentation from a different perspective. Rather than designing more lightweight networks with static structure, we propose a highly efficient framework with **d**ynamic **a**rchitecture for **med**ical volumetric segmentation (Med-DANet). As illustrated in Fig. 1 (b), taking a 2D image slice as input data, the Decision Network firstly generates a slice-dependent choice which represents the level of segmentation difficulty for the current slice. Then, according to the optimal choice made by the Decision Network, our method can adaptively determine to skip the current slice (i.e. directly output the segmentation map with only zero – background class) as highlighted with green boxes in Fig. 1 (b) or utilize the corresponding candidate segmentation network in the pre-defined Model Bank to accurately segment the current slice. The Model Bank consists of several networks with different model complexities. In this way, a reasonable allocation of computing resources for each slice is achieved by our adaptive segmentation framework. The main contributions of this work can be summarized as follows:

– This work presents the *first attempt* to explore the potential of dynamic inference in medical volumetric segmentation task. We focus on the 3D MRI

brain tumor segmentation and propose a new framework with dynamic architectures to achieve a good balance between segmentation accuracy and efficiency. The proposed Med-DANet is generic and can be applied to any volumetric segmentation tasks (see Supplementary for the experiments of our Med-DANet on liver tumor segmentation with CT images).
– By exploiting the special characteristics of multi-modal MRI brain tumor segmentation data that different slices have diverse degree of difficulty for segmentation, a comprehensive choice metric is designed to acquire the supervision signal for Decision Network, achieving the trade-off between accuracy and computational complexity of the model.
– Our proposed Med-DANet has strong scalability and flexibility. Any 2D networks can be incorporated into the Model Bank to meet various accuracy and efficiency requirements.
– Extensive experiments on two benchmark datasets (BraTS 2019 and BraTS 2020) for multi-modal 3D MRI brain tumor segmentation demonstrate that our method reaches competitive or better performance than previous state-of-the-art methods with much less model complexity.

## 2 Related Work

### 2.1 Static and Lightweight CNNs for Medical Image Segmentation

For medical image segmentation task, U-Net [27] and its variants [8,24,37] have achieved great success recently. However, the expensive computational costs impede the timely segmentation for assisting clinical diagnosis. To this end, great efforts have been made to design lightweight networks with improved model efficiency. For example, 3D-ESPNet [23] generalizes the efficient ESPNet [20] for 2D semantic segmentation to 3D medical volumetric segmentation, achieving satisfactory results on medical images. S3D-UNet [7] takes advantages of the separable 3D convolution to improve model efficiency. DMFNet [4] develops a novel 3D dilated multi-fiber network to bridge the gap between model efficiency and accuracy for 3D MRI brain tumor segmentation. HDCNet [18] replaces 3D convolutions with a novel hierarchical decoupled convolution (HDC) module to achieve a light-weight but efficient pseudo-3D model. [16] introduces a lightweight 3D U-Net with depth-wise separable convolution (DSC), which can not only avoid over fitting but also improve the generalization ability. In addition, knowledge distillation is also a popular method to achieve lightweight networks (i.e. student network). For example, [26] proposes an efficient architecture by distilling knowledge from well-trained medical image segmentation networks to train another lightweight network for efficient medical image segmentation.

### 2.2 Dynamic Networks for Efficient Inference

Lightweight models operates on the input data with the same static architecture, which cannot adaptively achieve the trade-off between accuracy and computational cost. To cope with this problem, dynamic networks are developed for

efficient and adaptive inference [31–35,38]. From the perspective of model architecture, the dynamic structure of network includes dynamic depth and dynamic width. For instance, slimmable networks [35] dynamically adjust the network width to achieve accuracy and efficient trade-offs at inference time. Moreover, adjusting input resolution is also an effective way to balance between accuracy and efficiency. DRNet [38] presents a novel dynamic-resolution network in which the resolution is determined dynamically based on each input sample.

Apart from the research on dynamic inference mostly for the classification task, a few works aim to achieve dynamic inference in pixel labeling tasks. Kong et al. [13] propose Pixel-wise Attention Gating to selectively process each pixel, allocating more computing power to pixels of fuzzy targets under specific resource constraints. Dynamic Multi-scale Network (DMN) [11] adaptively learns weights of convolution kernels according to different input instances, arranging multiple DMN branches to learn multi-scale semantic information in parallel. Li et al. [15] introduce the concept of dynamic routing to generate data-dependent routes. Based on the scale distribution of objects in an image, the proposed soft condition gates can adaptively select scale transformation routes in an end-to-end manner.

## 3 Methodology

### 3.1 Overview

An overview of our Med-DANet is shown in Fig. 2. In general, our framework consists of an extremely lightweight Decision Network ($\mathcal{D}$) and a Model Bank ($\mathcal{B}$) which contains $n$ different medical image segmentation networks ($M_1, M_2..., M_n$). Models in the bank should be diverse in terms of the number of parameters and computational cost. To deal with the medical image datasets where segmentation targets are sparsely distributed among slices, we learn a *slice-specific* decision by the Decision Network to dynamically select a suitable model from the Model Bank for the subsequent segmentation task, as formulated by Eq. 1.

$$y = \mathcal{D}(x) \circ \mathcal{B}(x), \tag{1}$$

where $x$ denotes the input image and $y$ is the corresponding prediction. $\mathcal{D} \circ \mathcal{B}$ indicates to take the matched element with index $\mathcal{D}$ in the collection $\mathcal{B}$, and the calculation details will be explained in the next subsection.

Roughly speaking, the Decision Network will comprehensively considers the segmentation accuracy and efficiency of each model, making the most appropriate choice. As for the Model Bank, any 2D networks can be included to meet various accuracy and efficiency requirements. More discussions on the model choices and ablation study are presented in Sec. 4.1 and Sec. 4.3.

### 3.2 Dynamic Selection Policy

We reduce the channel size of ShuffleNetV2 [19] to get an extremely lightweight classification network as our Decision Network so that its computational overhead is negligible in the entire framework. The Decision Network undertakes

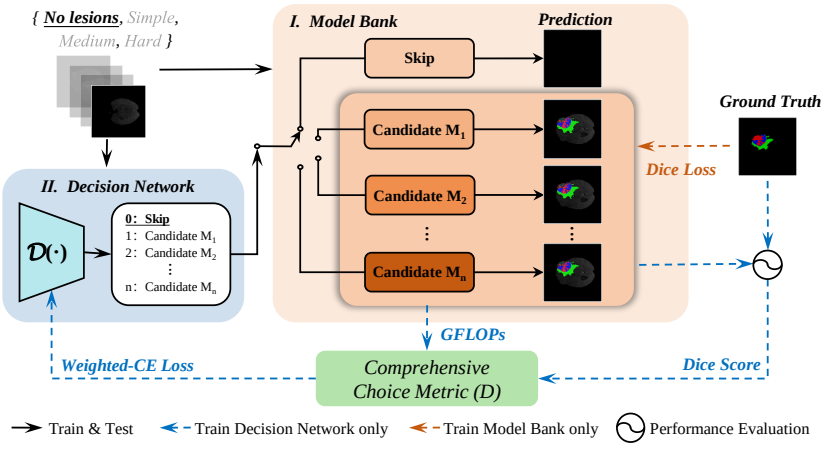



**Fig. 2.** The illustration of the overall architecture of our proposed Med-DANet. Taking a 2D image slice as input data, the Decision Network generates a slice-dependent choice which represents the level of segmentation difficulties for the current slice. Then, according to the optimal choice made by the Decision Network, our method can adaptively determine to skip the current slice (i.e. directly output a segmentation map with only zero – "background" class) or utilize the corresponding candidate network in the pre-defined Model Bank (containing several networks with different model complexities) to accurately segment the current slice.

a $n+1$-class classification task, and the $n+1$ categories refer to the $n$ candidate networks and a skip procedure. Therefore, the Decision Network and Model Bank can be respectively formulated as

$$\mathcal{D}(x) = \{\hat{D}|x;\theta\}, \tag{2}$$

$$\mathcal{B}(x) = [\varnothing, M_1(x), M_2(x), ..., M_n(x)]. \tag{3}$$

$\theta$ represents all the parameters of the Decision Network and $\hat{D}$ is the prediction of $\mathcal{D}(x)$. $M_1 \sim M_n$ denote the model candidates and $\varnothing$ indicates the skip procedure.

To be specific, when encountering a slice with only background (background slice, lesions are considered as foreground), the Decision Network will choose to directly skip the subsequent segmentation process. Otherwise, the Decision Network dynamically selects an appropriate segmentation model considering the recognition difficulty of foreground objects. During training process, the supervision of the Decision Network comes from the trade-off between model performance and efficiency. The calculation of the supervision signal of the decision process is as follows

$$D = \begin{cases} 0, & P_f < 1 \\ argmax((1-\alpha) * S_i + \alpha * softmax(\frac{1}{F_i})) + 1, & P_f \geqslant 1 \end{cases}, \tag{4}$$

where $S_i$ and $F_i$ is respectively the Dice Score and FLOPs of candidate model $M_i$ during the model training. $P_f$ denotes the number of foreground pixels (all pixels of segmentation targets). Specifically, if the number of foreground pixels is less than 1 (i.e. $P_f < 1$), the current slice will be considered without any lesion areas, which should be directly skipped (i.e. the corresponding supervision is 0) during inference. Note that we normalize $\frac{1}{F_i}$ through the softmax operation to avoid the negative effects of the order of magnitude difference between accuracy and computations, in case that the acquired $D$ is dominated by either model performance or complexity. In addition, $\alpha$ is a coefficient to moderate the impact of Dice Score and FLOPs.

Given the choice $\hat{D}$ predicted by $\mathcal{D}(x)$ and the ground-truth $D$ calculated with Eq. 4, we apply the weighted cross-entropy loss to supervise our Decision Network. This allows the network to learn to skip (assigning all pixels directly to the background class without going through the segmentation models) the pure background slices in the dataset and comprehensively measure the accuracy ($S_i$) as well as efficiency ($\frac{1}{F_i}$) of different models for the segmentation targets from individual slice. In practice, the skip procedure is essential and can be widely applied because some background slices barely capturing any image content are very common in medical volumes, it is pointless to invest too much computation on these background slices. Moreover, the recognition difficulty of segmentation targets varies from slice to slice, it is more efficient to dynamically select segmentation models of different complexities.

### 3.3 Training and Inference Strategy

**Training**. The training process of the entire framework consists of two steps. First, the ensemble training of segmentation models. To save the training time cost, the $n$ segmentation models are jointly trained, minimizing the mean average of the dice-losses of all models and performing gradient back-propagation synchronously.

$$diceloss_j = \sum_{i=1}^{C}(1 - \frac{2|pred_i \cap truth_i|}{|pred_i| + |truth_i|}), \tag{5}$$

$$Loss_{\mathcal{B}} = \frac{1}{n}\sum_{j=1}^{n} diceloss_j. \tag{6}$$

Here $C$ denotes the number of segmentation classes of the dataset, $diceloss_j$ is the dice-loss of candidate segmentation model $M_j (j \in \{1, 2, ..., n\})$ and $Loss_{\mathcal{B}}$ is the overall loss when training the Model Bank.

After that, we train the Decision Network with a weighted cross-entropy loss:

$$Loss_{\mathcal{D}} = WCE(D, \hat{D}) = -\sum_{i=0}^{n} w_i * d_i * log(\hat{d}_i), \tag{7}$$

where WCE is short for weighted cross-entropy, $d_i$ and $\hat{d}_i$ are respectively the ground-truth and logits predicted by the Decision Network for model candidate $i$, $w_i$ represents the corresponding loss weight.

To cope with the problem of class imbalance (background slices make up a considerable portion of the dataset) and further pursue a better trade-off between segmentation accuracy and model complexity, we slightly enlarge the loss weights of candidate models with better performance (i.e. relatively lower the loss weight of the skip procedure).

**Inference**. After the two-step training phase mentioned above, the well-trained decision network and predefined Model Bank are cascaded sequentially to achieve the final model structure at inference stage. Given a 2D slice as input image, our extremely lightweight Decision Network will decide to skip the current slice or choose the most appropriate segmentation network in the Model Bank based on the segmentation difficulty of the current slice. Following the specific selective choice made by Decision Network, the current slice will be directly skipped (i.e. output the corresponding segmentation maps with all zeros) or segmented by the single activated segmentation network included in the Model Bank. In this way, a dynamic slice-dependent framework with greatly improved efficiency is realized by our method. On one hand, compared with the previously proposed lightweight networks with static structure, our Med-DANet makes dynamic structure adjustments for different inputs instead of treating all inputs equally. On the other hand, compared with the previously proposed dynamic methods that utilize prediction confidence to determine whether the cascaded architecture need to early exit or not, our highly efficient Med-DANet can achieve the accurate segmentation in a one-pass manner.

## 4 Experiments

### 4.1 Experimental Setup

**Data and Evaluation Metric.** The first 3D MRI dataset used in the experiments is provided by the Brain Tumor Segmentation (BraTS) 2019 challenge [1, 2, 21]. It comprises 335 patient cases for training and 125 cases for validation. Each sample is composed of 3D brain MRI scans with four modalities. Each modality has a volume of $240 \times 240 \times 155$ that has already been aligned into the same space. The ground truth include 4 classes: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2) and GD-enhancing tumor (label 4). The segmentation accuracy is measured by the Dice score and the Hausdorff distance (95%) metrics for enhancing tumor region (ET, label 4), regions of the tumor core (TC, labels 1 and 4), and the whole tumor region (WT, labels 1,2 and 4), while the computational complexity is evaluated by the FLOPs metric. The second 3D MRI dataset is provided by the Brain Tumor Segmentation Challenge (BraTS) 2020 [1, 2, 21]. It is comprised of 369 cases for training, 125 cases for validation. Except for the number of samples in the dataset, the other information about these two MRI datasets are identical.

**Implementation Details.** The proposed Med-DANet is implemented on Pytorch [25] and trained with 2 NVIDIA Geforce RTX 3090 GPUs (each has 24GB memory). For the **training** aspect, we first jointly train the Model Bank for 400

epochs from scratch with a batch size of 64. To prevent the small-scale candidates in the Model Bank from overfitting and make sure the large models can be fully optimized, we let small-scale candidates detach the training process at the epoch of 300 and make large-scale candidates continue to back propagate in the remaining epochs. After acquiring the training labels for Decision Network using our proposed comprehensive choice metric, the Decision Network is trained for 50 epochs from scratch with a batch size of 64. The Adam optimizer and the poly learning rate strategy with warm-up are utilized to train both two parts of our method. The initial learning rate for training the Decision Network and Model Bank are 0.01 and 0.0001, respectively. Random cropping, random mirror flipping and random intensity shift are applied as the data augmentation techniques for training both the Decision Network and the segmentation candidates. The softmax Dice loss and weighted cross-entropy loss are employed to train the Model Bank and the Decision Network respectively. Besides, $L2$ Norm is applied for model regularization with a weight decay rate of $10^{-5}$.

As for the aspect of **model candidate selection** in the Model Bank, we choose the modified 2D UNet with various channel sizes (i.e. model width) and the 2D version of TransBTS [30] with different scales (i.e. model depth) in this paper. The reason of choosing these two baselines is that both of them are state-of-the-art methods for brain tumor segmentation with excellent performance and they also represent two popular network architectures (i.e. CNN and vision transformer) that can extract complementary information from the data. Compared with the original UNet [27], the modified version make improvements on both segmentation accuracy and efficiency. Taking consideration of both model depth and width, modified 2D UNet with a base channel of 12 (i.e. M1), modified 2D UNet with a base channel of 16 (i.e. M2), the light version of 2D TransBTS with 1-layer Transformer (i.e. M3), and 2D TransBTS with the original 4-layer Transformer (i.e. M4) are selected as the 4 model candidates in the Model Bank. According to the policy made by the Decision Network, the modified 2D UNet can segment the easy slice with greatly reduced computations, while the 2D version of TransBTS achieves precise segmentation of the difficult slices by modeling explicit long-range dependency. In this way, with the well-trained Decision Network and the splendid segmentation candidates in Model Bank, our framework can achieve great trade-off between segmentation accuracy and efficiency.

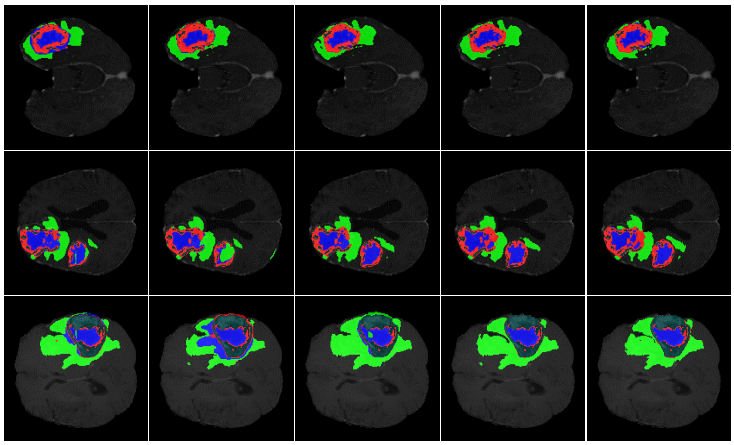
### 4.2 Results and Analysis

**BraTS 2019.** We conduct experiments on the BraTS 2019 validation set and compare our Med-DANet with previous state-of-the-art (SOTA) approaches.

The **quantitative results** are presented in Table 1. Our method achieves the Dice scores of 79.99%, 90.13%, 80.83% on ET, WT, TC, respectively, which are comparable or higher results than previous SOTA methods presented in Table 1. Besides, a considerable improvement has also been achieved for segmentation in terms of Hausdorff distance metric. It is worth noting that the model complexity of our Med-DANet is significantly less than other SOTA methods, while the segmentation performance of ours is extraordinary. For example, the

**Table 1.** Performance comparison on BraTS 2019 validation set. Per case and per slice denote the computational costs of segmenting a 3D case and a 2D slice separately.

| Method | Dice Score (%) ↑ | | | Hausdorff Dist. (mm) ↓ | | | FLOPs (G) ↓ | |
|---|---|---|---|---|---|---|---|---|
| | ET | WT | TC | ET | WT | TC | per case | per slice |
| 3D U-Net [8] | 70.86 | 87.38 | 72.48 | 5.062 | 9.432 | 8.719 | 1,669.53 | 13.04 |
| V-Net [22] | 73.89 | 88.73 | 76.56 | 6.131 | 6.256 | 8.705 | 749.29 | 5.85 |
| Attention U-Net [24] | 75.96 | 88.81 | 77.20 | 5.202 | 7.756 | 8.258 | 132.67 | 1.04 |
| Wang et al. [29] | 73.70 | 89.40 | 80.70 | 5.994 | 5.677 | 7.357 | - | - |
| Chen et al. [6] | 74.16 | **90.26** | 79.25 | 4.575 | **4.378** | 7.954 | - | - |
| Li et al. [14] | 77.10 | 88.60 | 81.30 | 6.033 | 6.232 | 7.409 | - | - |
| Frey et al. [9] | 78.70 | 89.60 | 80.00 | 6.005 | 8.171 | 8.241 | - | - |
| TransUNet [5] | 78.17 | 89.48 | 78.91 | 4.832 | 6.667 | 7.365 | 1205.76 | 9.42 |
| Swin-UNet [3] | 78.49 | 89.38 | 78.75 | 6.925 | 7.505 | 9.260 | 250.88 | 1.96 |
| TransBTS [30] | 78.36 | 88.89 | **81.41** | 5.908 | 7.599 | 7.584 | 333.09 | 2.60 |
| **Ours** | **79.99** | 90.13 | 80.83 | **4.086** | 5.826 | **6.886** | **77.78** | **0.61** |



3D U-Net VNet Att. U-Net **Ours** Ground Truth

**Fig. 3.** The visual comparison of MRI brain tumor segmentation results. The blue regions denote the enhancing tumors, the red regions denote the non-enhancing tumors, and the green ones denote the peritumoral edema.

computational complexity of TransBTS [30] is **3.5** times that of the proposed Med-DANet, and the model computational complexity of TransUNet [5] is surprisingly **15.4** times that of our method, which fully validates the effectiveness of adaptive architecture for dynamic inference.

For **qualitative analysis**, the brain tumor segmentation results of various methods are shown in Fig. 3 for a visual comparison (more visual comparison on BraTS 2019 dataset can be seen in Supplementary), including 3D U-Net [8], V-Net [22], Attention U-Net [24], and our Med-DANet. Since the labels for the validation set are not available, the five-fold cross-validation evaluation is conducted on the training set for all methods. It is obvious from Fig. 3 that our framework can delineate the brain tumors more accurately and generate much better segmentation masks with the powerful candidates as our dynamic options.

Since we successfully take advantage of both CNNs and Transformer for different inputs, both local details and global context can be captured by our method to achieve accurate segmentation of tumors.

**BraTS 2020.** We also evaluate our Med-DANet on BraTS 2020 validation set and the segmentation results are reported in Table 2. With the hyper-parameters on BraTS 2019 directly adopted for model training, our method achieves Dice scores of 80.57%, 90.28%, 81.34% and HD of 6.474mm, 6.718mm, 7.416mm on ET, WT, TC. Considerable gain has been made by our method in terms of ET. Besides, compared with 3D U-Net [8], V-Net [22] and Residual 3D U-Net, our method makes great improvements in both metrics. It is clear that our method not only shows significant superiority in model performance but also has the great advantage of computational efficiency, which reveals the benefit of leveraging dynamic inference for medical volumetric segmentation task.

**Table 2.** Performance comparison on BraTS 2020 validation set. Per case and per slice denote the computational costs of segmenting a 3D case and a 2D slice separately.

| Method | Dice Score (%) ↑ | | | Hausdorff Dist. (mm) ↓ | | | FLOPs (G) ↓ | |
|---|---|---|---|---|---|---|---|---|
| | ET | WT | TC | ET | WT | TC | per case | per slice |
| 3D U-Net [8] | 68.76 | 84.11 | 79.06 | 50.983 | 13.366 | 13.607 | 1,669.53 | 13.04 |
| V-Net [22] | 61.79 | 84.63 | 75.26 | 47.702 | 20.407 | 12.175 | 749.29 | 5.85 |
| Deeper V-Net [22] | 68.97 | 86.11 | 77.90 | 43.518 | 14.499 | 16.153 | - | - |
| 3D Residual U-Net [36] | 71.63 | 82.46 | 76.47 | 37.422 | 12.337 | 13.105 | 407.37 | 3.18 |
| Liu et al. [17] | 76.37 | 88.23 | 80.12 | 21.390 | 6.680 | **6.490** | - | - |
| Vu et al. [28] | 77.17 | **90.55** | **82.67** | 27.040 | **4.990** | 8.630 | - | - |
| Ghaffari et al. [10] | 78.00 | 90.00 | 82.00 | - | - | - | - | - |
| TransUNet [5] | 78.42 | 89.46 | 78.37 | 12.851 | 5.968 | 12.840 | 1205.76 | 9.42 |
| Swin-UNet [3] | 78.95 | 89.34 | 77.60 | **11.005** | 7.855 | 14.594 | 250.88 | 1.96 |
| TransBTS [30] | 78.50 | 89.00 | 81.36 | 16.716 | 6.469 | 10.468 | 333.09 | 2.60 |
| **Ours** | **80.57** | 90.28 | 81.34 | **6.474** | 6.718 | 7.416 | **77.71** | **0.61** |

### 4.3 Ablation Studies

We conduct extensive ablation experiments to verify the effectiveness of our framework and justify the rationale of its design choices based on five-fold cross-validation evaluations on the BraTS 2019 training set. (1) We make a fair comparison with each single candidate in the predefined Model Bank in terms of segmentation performance and computational cost. (2) We investigate the effect of different designs for the final choice metric, which stands for the acquired supervision signal for Decision Network to help the proposed framework achieve optimal trade-off between accuracy and efficiency. (3) We explore the effect of different lightweight networks for our Decision Network. (4) We also analyze the effect of different numbers of candidate networks in the Model Bank. Besides, please check Supplementary for more ablation study on BraTS 2019 training set.

**Comparison with Each Single Candidate in Model Bank.** We first compare our Med-DANet with all the candidates in Model Bank to demonstrate the powerful potential of dynamic architecture for medical volumetric segmentation. It is worth noting that the comparison is made under two different common

settings to comprehensively evaluate the proposed framework. The first setting is utilize the cropped image with a spatial resolution of 128×128 for training process and use slide-window technique to inference on original input with the spatial resolution of 240×240 (i.e. full resolution), while the second setting is to utilize the full resolution for both training and inference. As presented in Table 3, considerable improvements are achieved by our method in terms of both segmentation accuracy and model efficiency. Compared with the candidate M4 which has the largest model complexity under setting 1, our method achieves comparable performance with up to **8x** less computational costs. The same situation can be clearly seen under setting 2 in Table 3. With much less model complexity and great segmentation performance, our proposed Med-DANet pursues the best trade-off between accuracy and efficiency, demonstrating the significance of adaptive architecture for dynamic inference.

**Table 3.** Comparison with each single candidate in the predefined Model Bank.

| **Under Setting 1 (w/ slide window)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Method | Dice Score (%) ↑ | | | FLOPs (G) ↓ | | | |
| | ET | WT | TC | All Cases | Per Case | Per Slice | Per Inference |
| M1(Modified 2D UNet-12) | 75.49 | 90.21 | 81.58 | **99,026.40** | **1,500.40** | **9.68** | **0.61** |
| M2(Modified 2D UNet-16) | 78.31 | 90.61 | 82.59 | 174,646.57 | 2,646.16 | 17.07 | 1.07 |
| M3(2D TransBTS-light) | 78.72 | 90.30 | 82.99 | 385,957.45 | 5,847.84 | 37.73 | 2.36 |
| M4(2D TransBTS) | 77.21 | **91.08** | **83.27** | 814,962.73 | 12,347.92 | 79.66 | 4.98 |
| **Ours** | **78.75** | 90.40 | 83.13 | 102,398.01 | 1,551.485 | 10.01 | 0.63 |
| **Under Setting 2 (w/o slide window)** | | | | | | | |
| Method | Dice Score (%) ↑ | | | FLOPs (G) ↓ | | | |
| | ET | WT | TC | All Cases | Per Case | Per Slice | Per Inference |
| M1(Modified 2D UNet-12) | 76.86 | 90.16 | 80.43 | **21,748.98** | **329.53** | **2.13** | **2.13** |
| M2(Modified 2D UNet-16) | 77.22 | 89.99 | 81.61 | 38,362.50 | 581.25 | 3.75 | 3.75 |
| M3(2D TransBTS-light) | 76.64 | 90.16 | 82.21 | 90,862.86 | 1,376.71 | 8.88 | 8.88 |
| M4(2D TransBTS) | 76.48 | 90.57 | **83.21** | 203,351.94 | 3,081.09 | 19.88 | 19.88 |
| **Ours** | **77.73** | **90.65** | 82.72 | 38,211.12 | 578.96 | 3.74 | 3.74 |

**Effect of Different Designs for the Comprehensive Choice Metric.** To seek the optimal trade-off between model complexity and performance, we further investigate the effect of different designs for our proposed comprehensive choice metric (as illustrated in Eq. 4). As described in Sec. 3, we introduce $\alpha$ and softmax operation to moderate the impact of Dice Score and FLOPs, in case that the acquired ground truth for Decision Network is dominated by either accuracy or complexity. The ablation results are listed in Table 4. It shows that $\alpha = 0.001$ is the sweet spot for the whole framework to achieve the optimal balance between accuracy and efficiency. Specifically, increasing $\alpha$ will make our method focus more on model efficiency, while the decreasing of $\alpha$ will push our method to pursue model accuracy without the consideration of computational cost. Similarly, the drop of softmax operation on either Dice Scores or FLOPs will cause our framework to extremely pursue either the model performance or efficiency. By adopting the optimal configuration ($\alpha = 0.001$, with softmax on

FLOPs), our Med-DANet achieves greatly reduced computational complexity and competitive model accuracy.

**Table 4.** Ablation study on effect of different design for the proposed comprehensive choice metric. "*S*", "*F*" denote the Dice Scores and FLOPs respectively, w/o and w/ denote with or without softmax on corresponding metrics (i.e. Dice Scores and FLOPs).

| Comprehensive Choice Metric Design | Dice Score (%) ↑ | | | FLOPs (G) ↓ | | | |
|---|---|---|---|---|---|---|---|
| | ET | WT | TC | All Cases | Per Case | Per Slice | Per Inference |
| $\alpha = 0.0001$ | **78.84** | **90.48** | **83.28** | 129,768.88 | 1,966.20 | 12.69 | 0.79 |
| $\alpha = 0.001$ | 77.00 | 90.17 | 82.34 | 51,994.98 | 787.80 | 5.08 | 0.32 |
| $\alpha = 0.01$ | 77.21 | 90.19 | 81.80 | 49,951.03 | 756.83 | 4.88 | 0.31 |
| $\alpha = 0.001$, w/ S & F | 77.00 | 90.17 | 82.34 | 51,994.98 | 787.80 | 5.08 | 0.32 |
| $\alpha = 0.001$, w/o F | 75.57 | 90.10 | 81.85 | **48,255.83** | **731.15** | **4.72** | **0.29** |
| **$\alpha = 0.001$, w/o S** | 78.75 | 90.40 | 83.13 | 102,398.01 | 1,551.485 | 10.01 | 0.63 |
| $\alpha = 0.001$, w/o S & F | 77.12 | 90.31 | 82.66 | 69,618.70 | 1,054.83 | 6.81 | 0.43 |

**Effect of Different Lightweight Networks for Decision Network.** After investigating of the best design for the choice metric, we verify the effectiveness of our method with different Decision Networks. To achieve the highly efficient overall framework, the computational cost brought by the Decision Network should be controlled within acceptable limits. Therefore, four lightweight CNNs (MobileNetV2, GhostNet, ShuffleNetV2, and our modified ShuffleNetV2) are selected to study the influence of the Decision Network. To be noticed, the modified ShuffleNetV2 is acquired by greatly cutting down the channel size (i.e. model width). As shown in Table 5, with our modified ShuffleNetV2 as the Decision Network, our Med-DANet yields the best trade-off between accuracy and computational cost. Although our method achieves the best Dice Scores with MobileNetV2 as the Decision Network, the model complexity of MobileNetV2 and the overall FLOPs resulted by the guidance of MobileNetV2 are not acceptable. Specifically, the model complexity of the modified ShuffleNetV2 is approximately **1/3** of GhostNet or ShuffleNetV2 and nearly **1/23** of MobileNetV2, which shows the effectiveness and efficiency of our optimal Decision Network. It is clear that employing modified ShuffleNetV2 enables our framework to show great superiority in terms of computation with competitive model performance.

**Table 5.** Ablation study on effect of different choices for Decision Network. DN denotes the Decision Network, while ShuffleNetV2-M denotes our modified ShuffleNetV2.

| Decision Network | DN's FLOPs (G) | Dice Score (%) ↑ | | | Overall FLOPs (G) ↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | | ET | WT | TC | All Cases | Per Case | Per Slice | Per Inference |
| MobileNetV2 | 1.758 | **78.54** | **90.22** | **82.44** | 334,046.58 | 5,061.31 | 32.65 | 2.04 |
| GhostNet | 0.278 | 75.57 | 90.19 | 81.48 | 82,536.60 | 1,250.56 | 8.07 | 0.50 |
| ShuffleNetV2 | 0.247 | 77.20 | 90.19 | 81.48 | 96,606.43 | 1,463.73 | 9.44 | 0.59 |
| **ShuffleNetV2-M** | 0.078 | 77.00 | 90.17 | 82.34 | **51,994.98** | **787.80** | **5.08** | **0.32** |

**Effect of Different numbers of Candidate Networks in Model Bank.** Finally, we conduct experiments to investigate the influence of the number of can-

didates in Model Bank on segmentation performance and efficiency. The quantitative results are illustrated in Table 6. First of all, with no candidates (only skip procedure), the framework will naturally not work at all. Then we add the lightest CNN and Transformer as candidates ($n = 2$), a good result has been achieved already. After that, the largest CNN and Transformer are also incorporated to Model Bank ($n = 4$), making the segmentation performance and efficiency of the network both improve. Compared to 2 candidates, 4 candidates give the network more options for pursuing either performance or efficiency. However, when we further add 2 medium-sized CNN and Transformer to the Model Bank ($n = 6$), although the network performance (i.e. segmentation accuracy) is further improved because of more optional network candidates, the computational cost is also increased. Moreover, more candidate networks in the model bank would also increase the training cost. If higher precision requirements is necessary for the segmentation tasks, more candidates can be plugged into the Model bank to further boost the final performance. In conclusion, 4 candidates in Model Bank achieve the best balance between the accuracy and efficiency.

**Table 6.** Ablation study on effect of different numbers of candidate networks ($n$).

| number of candidate networks | Dice Score (%) ↑ | | | FLOPs (G) ↓ | | | |
|---|---|---|---|---|---|---|---|
| | ET | WT | TC | All Cases | Per Case | Per Slice | Per Inference |
| 0 (only skip) | 9.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2 (1 CNN + 1 TR) | 75.60 | 90.15 | 82.11 | 54,247.84 | 821.94 | 5.30 | 0.33 |
| **4 (2 CNN + 2 TR)** | 77.00 | 90.17 | 82.34 | **51,994.98** | **787.80** | **5.08** | **0.32** |
| 6 (3 CNN + 3 TR) | **78.71** | **90.76** | **82.92** | 84,130.02 | 1,274.70 | 8.22 | 0.51 |

## 5 Conclusion

We present the *first attempt* to explore the potential of dynamic inference in medical volumetric segmentation task. We focus on the 3D MRI brain tumor segmentation and propose a new framework named Med-DANet with dynamic architectures to achieve the trade-off between segmentation accuracy and efficiency. The proposed Med-DANet is generic and not limited to MRI brain tumor segmentation, which can be applied to any volumetric segmentation tasks. It is also worth noting that our proposed Med-DANet has strong scalability and flexibility. Any 2D state-of-the-art methods can be incorporated into our framework to satisfy different accuracy and efficiency requirements. Extensive experiments on two benchmark datasets (BraTS 2019 and BraTS 2020) for multi-modal 3D MRI brain tumor segmentation demonstrate that our Med-DANet reaches competitive or better performance than previous state-of-the-art methods with greatly improved model complexity.

**Acknowledgment** This work was supported by the Fundamental Research Funds for the China Central Universities of USTB (FRF-DF-19-002), Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (BK20BE014).

## References


1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**, 170117 (2017) 2, 8
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018) 2, 8
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021) 10, 11
4. Chen, C., Liu, X., Ding, M., Zheng, J., Li, J.: 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 184–192. Springer (2019) 2, 4
5. 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) 10, 11
6. Chen, M., Wu, Y., Wu, J.: Aggregating multi-scale prediction based on 3d u-net in brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 142–152. Springer (2019) 10
7. Chen, W., Liu, B., Peng, S., Sun, J., Qiao, X.: S3d-unet: separable 3d u-net for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 358–368. Springer (2018) 2, 4
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016) 2, 4, 10, 11
9. Frey, M., Nau, M.: Memory efficient brain tumor segmentation using an autoencoder-regularized u-net. In: International MICCAI Brainlesion Workshop. pp. 388–396. Springer (2019) 10
10. Ghaffari, M., Sowmya, A., Oliver, R.: Brain tumour segmentation using cascaded 3d densely-connected u-net. arXiv preprint arXiv:2009.07563 (2020) 11
11. He, J., Deng, Z., Qiao, Y.: Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3562–3572 (2019) 5
12. Huo, Y., Liu, J., Xu, Z., Harrigan, R.L., Assad, A., Abramson, R.G., Landman, B.A.: Robust multicontrast mri spleen segmentation for splenomegaly using multi-atlas segmentation. IEEE Transactions on Biomedical Engineering **65**(2), 336–343 (2017) 2
13. Kong, S., Fowlkes, C.: Pixel-wise attentional gating for scene parsing. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1024–1033. IEEE (2019) 5
14. Li, X., Luo, G., Wang, K.: Multi-step cascaded networks for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 163–173. Springer (2019) 10

15. Li, Y., Song, L., Chen, Y., Li, Z., Zhang, X., Wang, X., Sun, J.: Learning dynamic routing for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8553–8562 (2020) 5
16. Li, Z., Pan, J., Wu, H., Wen, Z., Qin, J.: Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3d u-net. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–206. Springer (2020) 2, 4
17. Liu, C., Ding, W., Li, L., Zhang, Z., Pei, C., Huang, L., Zhuang, X.: Brain tumor segmentation network using attention-based fusion and spatial relationship constraint. In: International MICCAI Brainlesion Workshop. pp. 219–229. Springer (2020) 11
18. Luo, Z., Jia, Z., Yuan, Z., Peng, J.: Hdc-net: Hierarchical decoupled convolution network for brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics **25**(3), 737–745 (2020) 2, 4
19. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018) 5
20. Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., Hajishirzi, H.: Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the european conference on computer vision (ECCV). pp. 552–568 (2018) 4
21. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014) 2, 8
22. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016) 2, 10, 11
23. Nuechterlein, N., Mehta, S.: 3d-espnet with pyramidal refinement for volumetric brain tumor image segmentation. In: International MICCAI Brainlesion Workshop. pp. 245–253. Springer (2018) 2, 4
24. 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. arXiv preprint arXiv:1804.03999 (2018) 2, 4, 10
25. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019) 8
26. Qin, D., Bu, J.J., Liu, Z., Shen, X., Zhou, S., Gu, J.J., Wang, Z.H., Wu, L., Dai, H.F.: Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging **40**(12), 3820–3831 (2021) 2, 4
27. 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) 2, 4, 9
28. Vu, M.H., Nyholm, T., Löfstedt, T.: Multi-decoder networks with multi-denoising inputs for tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 412–423. Springer (2020) 11
29. Wang, F., Jiang, R., Zheng, L., Meng, C., Biswal, B.: 3d u-net based brain tumor segmentation and survival days prediction. In: International MICCAI Brainlesion Workshop. pp. 131–141. Springer (2019) 10

30. 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. Springer (2021) 9, 10, 11
31. Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G.: Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length. arXiv e-prints pp. arXiv–2105 (2021) 5
32. Yang, L., Han, Y., Chen, X., Song, S., Dai, J., Huang, G.: Resolution adaptive networks for efficient inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2369–2378 (2020) 5
33. Yang, T., Zhu, S., Chen, C., Yan, S., Zhang, M., Willis, A.: Mutualnet: Adaptive convnet via mutual learning from network width and resolution. In: European conference on computer vision. pp. 299–315. Springer (2020) 5
34. Yang, T., Zhu, S., Mendieta, M., Wang, P., Balakrishnan, R., Lee, M., Han, T., Shah, M., Chen, C.: Mutualnet: Adaptive convnet via mutual learning from different model configurations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 5
35. Yu, J., Yang, L., Xu, N., Yang, J., Huang, T.: Slimmable neural networks. arXiv preprint arXiv:1812.08928 (2018) 5
36. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018) 2, 11
37. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018) 2, 4
38. Zhu, M., Han, K., Wu, E., Zhang, Q., Nie, Y., Lan, Z., Wang, Y.: Dynamic resolution network. Advances in Neural Information Processing Systems **34** (2021) 5