Part I: Experimental results added for performance comparison and computational cost based on reviewer's comments

Preface: Firstly, we will present the DSNet50 network architecture that we have developed. We then discuss its performance in semantic segmentation, specifically using the Synapse medical image dataset and the Cityscapes natural image dataset. Finally, we will showcase the performance of our DSNet50 in tasks including natural image classification and object detection using datasets like ImageNet and COCO.

In order to improve performance of semantic segmentation, we propose the DSNet50-A to mitigate the impact of the inter-class similarity issuer. As a result, DSNet50-A surpasses FcaNet50 in semantic segmentation in our experiments, while achieving comparable outcomes in classification and detection (slightly weaker in classification and detection compared to FcaNet50 but with higher "**Train FPS**" index). To achieve better classification and object detection results, we propose DSNet50-B, which has a similar architecture to FcaNet50 but fewer parameters than FcaNet50 and DSNet50-A. The performance of classification and object detection of DSNet50-B lies between FcaNet50 and DSNet50-A.

Table 1: The designed architecture in our proposed DSNet50 network. DSNet50-A is proposed to improve performance of semantic segmentation, while DSNet50-B is designed to achieve better classification and object detection results.

Method	**Backbone**	**Plugin of every stage**				**Position of plugins**	**Ratio**
FcaNet50	ResNet50	MSCA	MSCA	MSCA	MSCA*	after conv3	16
DSNet50-A	ResNet50	None	None	None	DSM**	after conv2	2
DSNet50-B	ResNet50	DSM	DSM	DSM	DSM	after conv1	16

- MSCA*: Multi-spectral channel attention
- DSM**: Decoupled self-attention module

Section A: Network performance of semantic segmentation on the Synapse medical image dataset and the Cityscapes natural image dataset.

1.1 Experimental results of semantic segmentation on the Synapse dataset

Table 2: Experimental results of semantic segmentation on the Synapse dataset. Our proposed DSNet50-A achieves better segmentation results.

Method	**Parameters**	**FLOPs**	**Train FPS**	**mDice**	**mIoU**
ResNet50	47.13M	197.86G	7.52*	82.92	74.72
FcaNet50	49.62M	196.67G	81.60	84.06	76.1
DSNet50-A(**ours**)	50.67M	205.93G	60.4	**85.07**	**77.61**
DSNet50-B(**ours**)	48.94M	204.44G	**95.72**	83.8	75.76

*Train on a single graphics card A5000, while train on 8 graphics cards 4090.

1.2 Experimental results of semantic segmentation on the Cityscapes dataset

Table 3: Experimental results of semantic segmentation on the Cityscapes dataset. Our proposed DSNet50-A achieves better segmentation results.

Decoder	**Method**	**Pre-trained weights***	**Parameters**	**FLOPs**	**Train FPS**	**mIoU**	**mAcc**	**aAcc**

FCN	ResNet50 V1C**	-	47.13M	395.76G	-	72.25***	-	-
FCN	FcaNet50	FcaNet50	49.65M	395.91G	48.24	75.63	82.93	95.78
FCN	DSNet50-A(**ours**)	DSNet50-A	50.68M	411.91G	46.00	**76.01**	**83.59**	**95.86**
FCN	DSNet50-B(**ours**)	DSNet50-B	48.95M	408.92G	**49.44**	75.25	83.00	95.62
FCN	ResNet50 V1C	ResNet50 V1C	47.13M	395.76G	45.6	75.51	83.19	95.89
FCN	FcaNet50	ResNet50 V1C	49.65M	395.91G	69.76	76.45	83.38	95.93
FCN	DSNet50-A	ResNet50 V1C	50.68M	411.91G	**70.96**	**77.25**	**84.48**	**96.83**

Note: Due to time constraints, we were unable to pre-train FcaNet50 and DSNet50 using ResNet V1C on ImageNet1K dataset. Instead, we employed two training strategies: (1) utilizing pre-training weights for FcaNet50 and DSNet50 based on ResNet (specifically, ResNet V1B); (2) Both models were trained using the weights of ResNet50 V1C.

**Compare to ResNet, ResNet V1C replace the 7x7 conv in the stem with three 3x3 convs.

***The mIoU results of ResNet50 are obtained using official results [3] from mmseg [2].

Section B: Natural image classification on ImageNet1K dataset and object detection on COCO2018 dataset.

2.1 Experimental results of image classification on the ImageNet1K dataset

The experiments in Table4 are all conducted in mmpretrain [1], so there is a difference from the FcaNet paper report (the FcaNet paper used the Nvidia APEX mixed precision training toolkit, but due to limitations in experimental conditions, we did not use mixed precision).

Table 4: Experimental results of image classification on the ImageNet1K dataset. Although the results of our DSNet50-B are slightly inferior to FcaNet50, our “**Train FPS**” index is significantly better, indicating that our method has a faster training speed. However, it is important to note that both our method and FcaNet50 achieve comparable results in terms of accuracy.

Method	Parameters	FLOPs	Train FPS	top1	top5
ResNet50	25.557M	4.109G	1677.51	76.48	93.17
FcaNet50	30.121M	4.112G	1036.70	**77.50**	**93.79**
DSNet50-A(**ours**)	27.369M	4.357G	**1873.95**	76.51	93.12
DSNet50-B(**ours**)	29.106M	4.208G	1424.30	77.28	93.60

2.2 Experimental results of object detection on the COCO2018 dataset

Our DSNet50-B has a smaller number of parameters compared to FcaNet. Additionally, the “**Train FPS**” index of DSNet50-B is more than twice that of FcaNet, which indicates that our network has achieved a faster training speed.

Table 5: Experimental results of object detection on the COCO2018 dataset. Although the results of our DSNet50-B are slightly inferior to FcaNet50, our “**Train FPS**” index is more than twice that of FcaNet, indicating that our method has a faster training speed.

Model	Method	Parameters	FLOPs	Train FPS	AP	AP_{50}	AP_{75}	AP_{S}	AP_{M}	AP_{L}
Faster-RCNN	ResNet50	41.750M	187.20G	46.75	37.4	58.3	40.5	21.9	40.7	48.1
Faster-RCNN	FcaNet50	44.268M	187.31G	28.90	**38.9**	**60.2**	**42.4**	**23.1**	**42.5**	**49.9**
Faster-RCNN	DSNet50-A	45.302M	188.97G	44.95	37.8	59.4	40.8	23	41.6	48.1
Faster-RCNN	DSNet50-B	43.565M	194.27G	**64.89**	38.2	59.6	41.5	22.8	42.1	48.5

Reference

[1] https://github.com/open-mmlab/mmpretrain

[2] https://github.com/open-mmlab/mmsegmentation

[3] https://github.com/open-mmlab/mmsegmentation/tree/main/configs/fcn

Part II: Experimental results added for computational time based on reviewer's comments

Preface: In the Supplementary Material, it has been observed that our DSNet50 converges with fewer iterative steps compared to the baseline ResNet50. Now, let us reconsider my method from the perspective of computational time.

Section A: Computational time of semantic segmentation on the Synapse medical image dataset.

Based on the results in Fig. 1, it is evident that DSNet50-B achieves convergence in the shortest amount of time, making it a superior option compared to the ResNet50 and FcaNet50. However, DSNet50-A demonstrates a convergence speed similar to that of ResNet50. We hypothesize that a low 'Ratio' (as shown in Table 1) leads to a higher feature dimension, which in turn increases computational time in dense semantic segmentation.

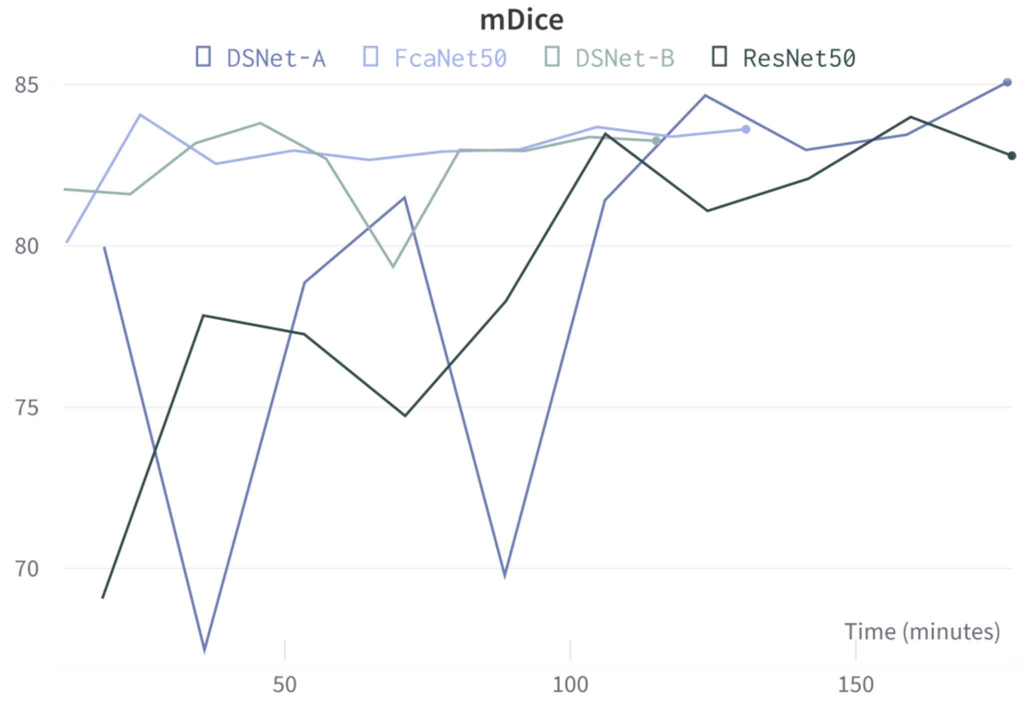

Fig. 1: Computational time of our proposed DSNet50 against compared networks on the Synapse dataset. The required computational time is ranked as follows: DSNet50-B < FcaNet50 < DSNet50-A < ResNet50.

Section B: Computational time of image classification on the ImageNet1K dataset.

Fig. 2 shows the required computational time is ranked as follows: DSNet50-A < ResNet50 < DSNet50-B < FcaNet50. Our proposed method, DSNet50-A, achieved optimal results in ImageNet1K dataset. Our both DSNet50-A and DSNet50-B demonstrate superior performance compared to FcaNet, indicating faster convergence performance.

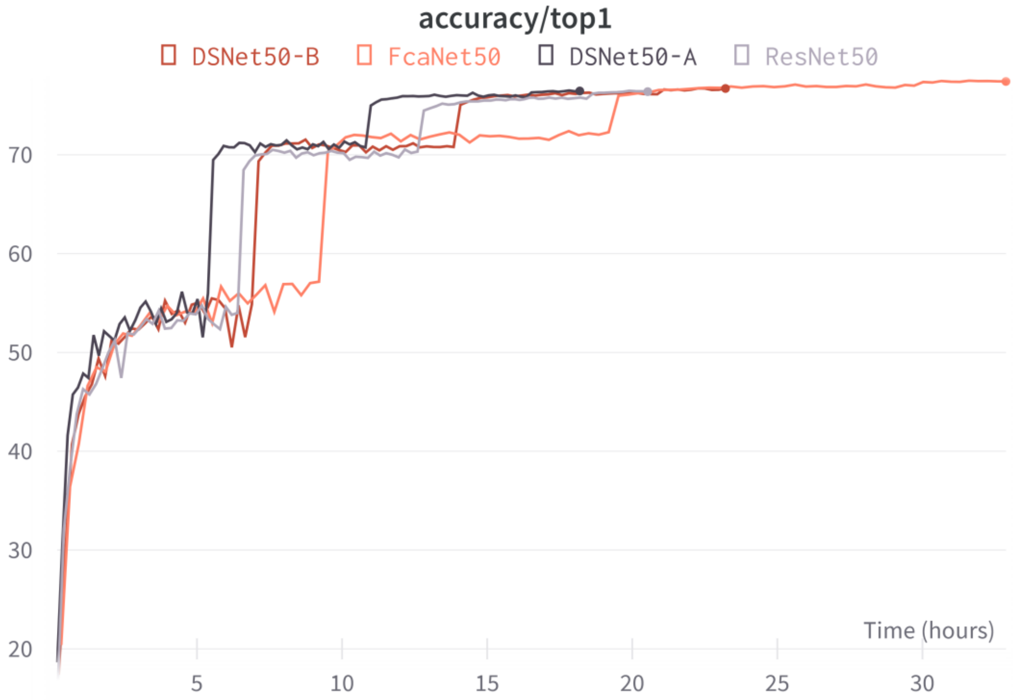

Fig 2: Computational time of our proposed DSNet50 against compared networks on the ImageNet1K dataset. The required computational time is ranked as follows: DSNet50-A < ResNet50 < DSNet50-B < FcaNet50.