DataFormer: Differential Additive Transformer for Lightweight Semantic Segmentation

Supplementary Material

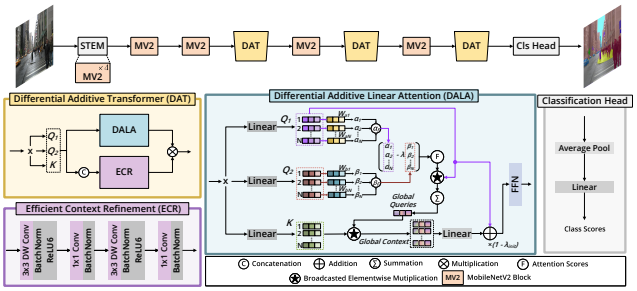

Figure A. The architecture of the proposed DataFormer model for the task of image classification.

This supplementary material presents the details of the ImageNet pre-training and classification results (Sec. 1), a comprehensive description of the model architecture (Sec. 2), the Feed-Forward Network's design (Sec. 3), as well as a discussion of the model's limitations and potential future work (Sec. 4).

1. ImageNet Pre-training

To ensure a fair basis for comparison, the DataFormer model was initialized using pre-trained parameters obtained from ImageNet. As depicted in Fig. A, the classification architecture of DataFormer consists of an average pooling layer, which is followed by a linear layer to effectively utilize global semantic features for generating class scores. Importantly, no Unified Feature Aggregation block is incorporated into the design. The model processes input images with a resolution of 224×224 to compute the classification scores. The quantitative performance of the DataFormer model on the ImageNet-1K dataset is presented in Tab. A.

Table A. DataFormer results for ImageNet classification.

Method	Input Size	Top-1 Accuracy(%)	GFLOPs	Parameters
DataFormer	224×224	66.5	0.1	1.68M

2. Detailed Network Structure

The detailed architecture of the DataFormer model, specifically developed for lightweight semantic segmentation, is outlined in Tab. B. In this context, N represents the total number of layers, while H corresponds to the number of heads. The input resolution considered for the model is 512×512.

Table B. Architectural details of DataFormer model. For input resolution of 512×512.

Stages	Details					Output Resolution
	layer	kernel size	expand ratio	output channels	stride	
Stage1	Conv	3	-	16	2	
	MobileNetV2	3	1	16	1	256×256
Stage2	MobileNetV2	3	4	16	2	
	MobileNetV2	3	3	16	1	128×128
Stage3	MobileNetV2	5	3	32	2	
	MobileNetV2	5	3	32	1	64×64
Stage4	MobileNetV2	3	3	64	2	
	MobileNetV2	3	3	64	1	32×32
Stage5	MobileNetV2	5	3	128	2	
	DAT Block N=2, H=4					16×16
Stage6	MobileNetV2	3	6	160	2	
	DAT Block N=2, H=4					8×8

3. Details of Feed-Forward Network

In the proposed DataFormer model, features extracted from the DALA block are passed to the Feed-Forward Network (FFN) for further processing. The FFN employs a 3×3 depth-wise convolution layer nestled between two 1×1 convolution layers enabling the network to effectively refine and capture robust semantic features. The design of the FFN is illustrated in Fig. B.

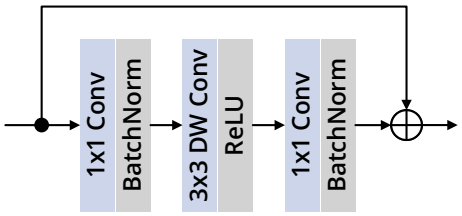

Figure B. Design of Feed-Forward Network.

4. Limitations and Future Work

The proposed model delivers strong performance and efficiency in semantic segmentation tasks, though certain challenges remain. One key limitation of lightweight models, including ours, is their reliance on ImageNet-1K pre-training to achieve optimal performance, as the absence of such pre-training significantly reduces models' effectiveness. While the proposed model has already demonstrated versatility through results on the object detection task, future work will focus on enhancing its robustness across diverse datasets and practical scenarios. Furthermore, optimizing the model for real-world applications will remain a priority to maximize its practical utility.