HiFormer: Hierarchical Multi-scale Representations Using transformers for Medical Image Segmentation
Supplementary Material

Moein Heidari*,1 Amirhossein Kazerouni*,1 Milad Soltany*,1 Reza Azad[2]
Ehsan Khodapanah Aghdam[3] Julien Cohen-Adad[4] Dorit Merhof †,5,6
[1] School of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran
[2] Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
[3] Department of Electrical Engineering, Shahid Beheshti University, Tehran, Iran
[4] MILA, Quebec AI Institute, Montreal, Canada
[5] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[6] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
moein_heidari@elec.iust.ac.ir, {amirhossein477, soltany.m.99, ehsan.khpaghdam}@gmail.com
azad@lfb.rwth-aachen.de, jcohen@polymtl.ca, dorit.merhof@ur.de

This supplementary material contains the following additional information. We expand Table 6 in the paper and provide more experiments regarding the impact of the proposed DLF module (see Table 1-2). Moreover, we provide additional information regarding the optimality of DLF module and the intuitions behind fusion of features in different levels and feature consistency.

A. Impact and justification of the DLF Module

A.1. Model Design Motivation

The deficiency of transformers in capturing local features, lack of data in the medical domain, and proven usage of CNN-produced features as an input to transformers and their success in vision tasks [2] led us to use a rich CNN backbone before the transformer. Subsequently, we used a successive Swin Transformer to capture multi-scale global dependencies. With respect to the deformation of body organs and tissues and diverse sizes and scales of neighboring organs, we proposed a representative token by applying GAP to form a representative token or messenger token-like [3, 4] to exchange information between scales to prevent the globality bias to a specific region and implicitly remembers its previous stage attention regions. Our expansion of ablation study for DLF module presence is depicted in Table 1 and Table 2.

In addition, the inspiration behind presenting variable HiFormer designs is to develop a general model with different scales and exhibit the stability of the model. Considering the accuracy-speed trade-off, one can exploit the S, L, or B HiFormer and investigate the whole network performance.

A.2. Hyper-parameter Optimization

S, L, r: We have considered the effect of network deepening and model scaling on the network performance similar to [2, 1]. In our experiments, we deduced that increasing the (S, L, r) pairs would lead to a substantial computational cost which is in contradiction to our main contribution of designing a stable model with low parameters. Specifically, we hypothesized that considering $S, L, r > 3$ (see row D in table 8) can result in the model overparameterization along with the extraction of redundant features. Hence, we adopted $(S, L, r) <= 3$.

Number of heads. Considering the number of heads of the transformer, we performed cross-validation using the synapse dataset and attained 6 heads as the ideal choice. For the sake of demonstrating the effect of the number of transformer heads, two more configurations besides 6, its half, and double (3 and 12) were considered in our ablation study.

B. Clarification on CNN backbones

Our study considered different backbones typically used in the literature [2], such as ResNet and DenseNet. Benefiting from the skip-connection criteria, the ResNet architecture can facilitate multi-level representation. Although a DensNet or ResNet with more layers might bring a stronger

*Equal contribution
†Corresponding author

Table 1. Impact of the DLF module on the skin lesion segmentation datasets.

Model	DLF	DSC	SE	SP	ACC
		ISIC 2017			
HiFormer-B	✗	0.9167	0.8814	**0.9895**	0.9678
HiFormer-B	✓	**0.9253**	**0.9155**	0.9840	**0.9702**
		ISIC 2018			
HiFormer-B	✗	0.8986	0.8559	**0.9870**	0.9595
HiFormer-B	✓	**0.9102**	**0.9119**	0.9755	**0.9621**
		PH^2			
HiFormer-B	✗	0.9321	0.9016	**0.9848**	0.9586
HiFormer-B	✓	**0.9460**	**0.9420**	0.9772	**0.9661**

Table 2. Impact of the DLF module on the *SegPC* dataset.

Model	DLF	mIoU
HiFormer-B	✗	0.9317
HiFormer-B	✓	**0.9406**

representation, it can be a more high-level representation so that its amalgamation with the transformer in subsequent layers would prevent the transformer from extracting better features. Moreover, the features attained from a ResNet with 50 layers can be considered as more general ones compared to 18, 34 layered shallow ResNets aiding the transformer in more optimal performance.

C. Feature Consistency

We provide two experiments to demystify the feature consistency. First, we present the feature visualization of each level before and after involving the DLF module (see Fig. 1-2). As illustrated, before the DLF module is applied, the attention location is more diffused, therefore the organ is not clearly emphasized. However, after applying the DLF module, attention is drawn to the desired organ and is more highlighted surrounding the organ, demonstrating that the DLF module makes features more consistent. Furthermore, both levels serve a complimentary function, with the larger level providing fine-grained features and the smaller level attempting to give extra information. As a result, both levels are required for the model to function effectively.

In the second experiment, we take the HiFormer-B and remove modules in a hierarchical order to observe how the features become consistent. As shown in Table 3, using only ResNet50 as the CNN module and dismissing others achieves a 77.40 dice score and 26.71 HD. Having involved the Swin Transformer, HD witnesses a 9.93 drop, indicating that our predictions become closer to their corresponding labels or more similar. Subsequently, applying the DLF module not only increases the dice score but also decreases HD, exhibiting that the module dramatically assists in making the features consistent.

Table 3. Impact of each module in HiFormer-B.

Model	CNN	transformer	DLF	DSC	HD
HiFormer-B	✓	✗	✗	77.40	26.71
HiFormer-B	✓	✓	✗	77.15	16.88
HiFormer-B	✓	✓	✓	**80.39**	**14.70**

References

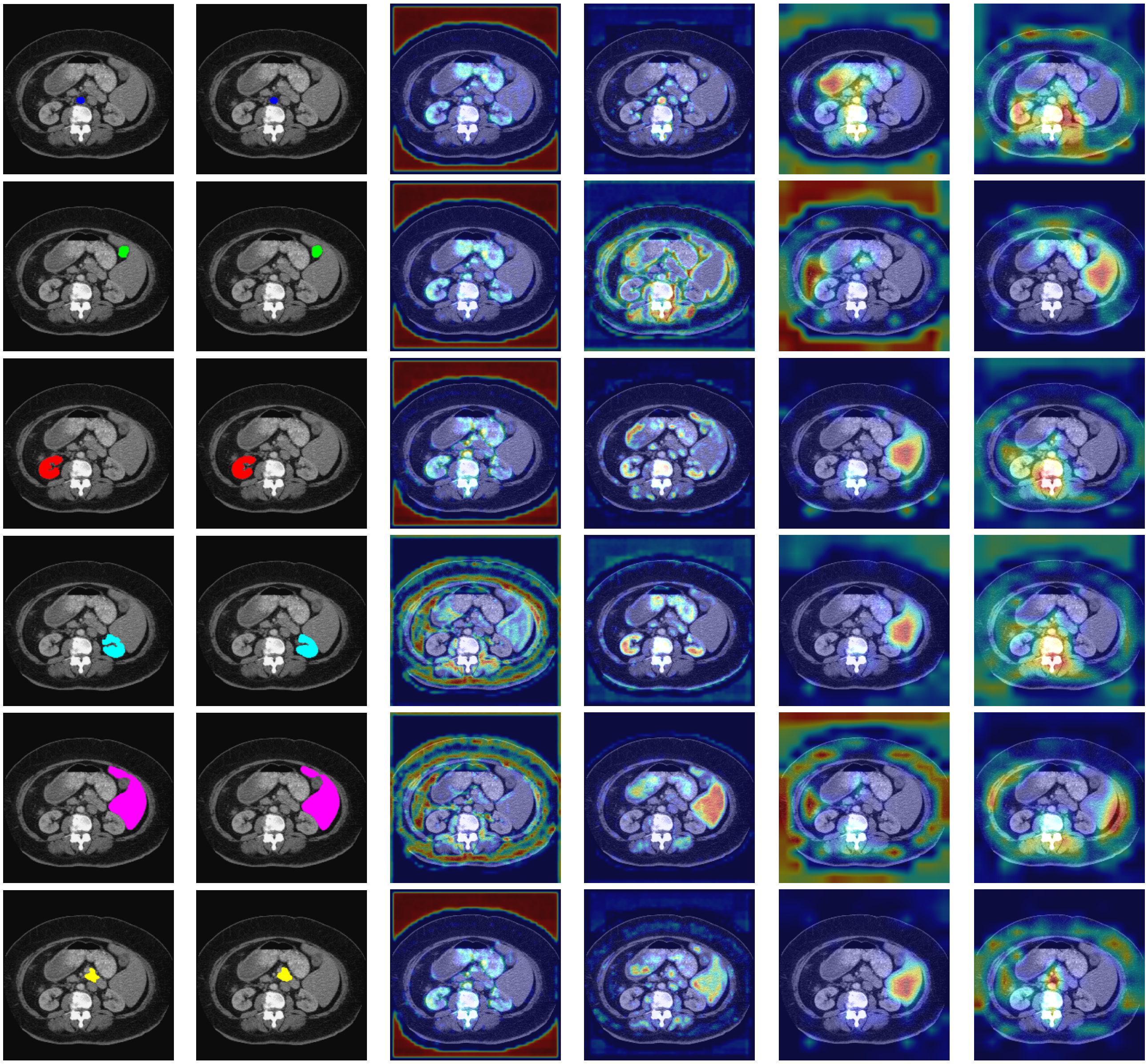
[1] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[3] Jiemin Fang, Lingxi Xie, Xinggang Wang, Xiaopeng Zhang, Wenyu Liu, and Qi Tian. Msg-transformer: Exchanging local spatial information by manipulating messenger tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12063–12072, 2022.

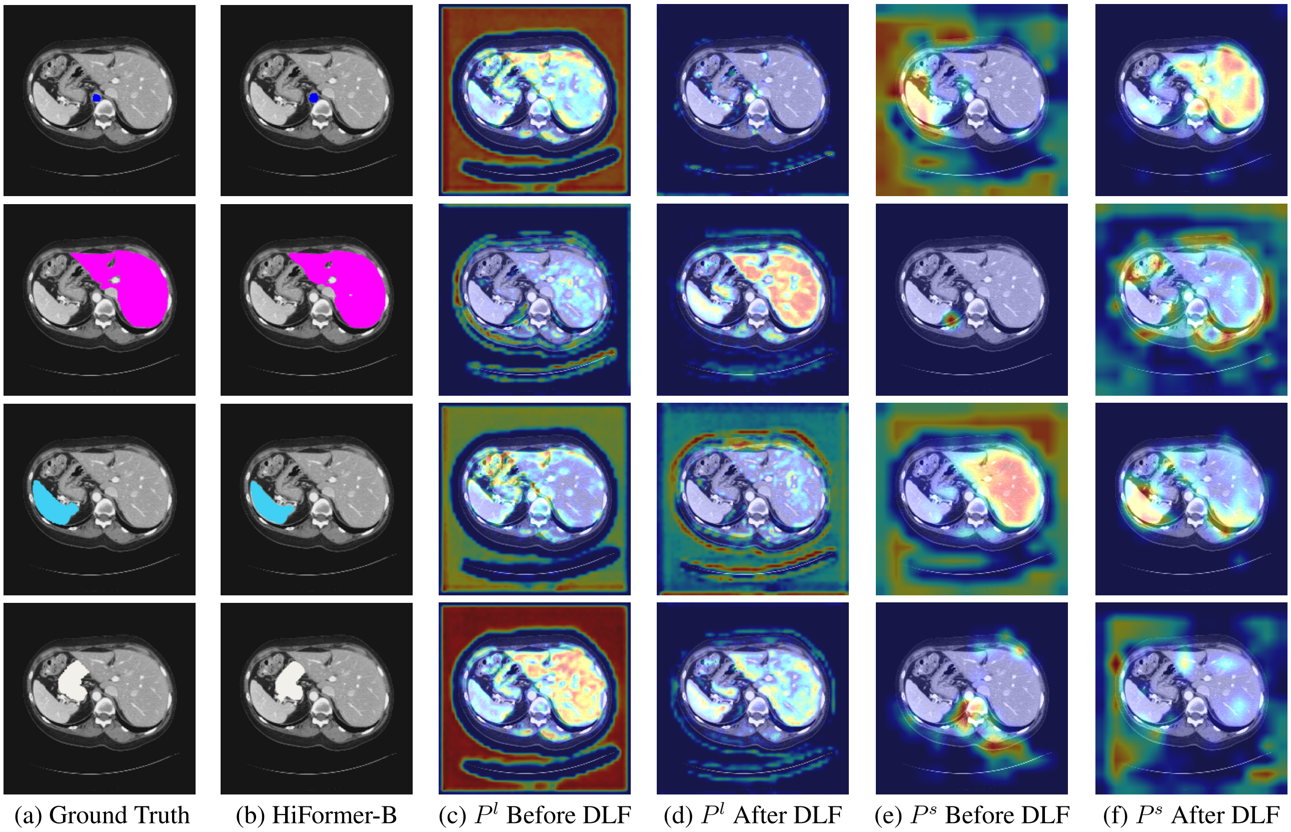
[4] Ammarah Farooq, Muhammad Awais, Sara Ahmed, and Josef Kittler. Global interaction modelling in vision transformer via super tokens. *arXiv preprint arXiv:2111.13156*, 2021.

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.



(a) Ground Truth (b) HiFormer-B (c) P^l Before DLF (d) P^l After DLF (e) P^s Before DLF (f) P^s After DLF

Figure 1. Feature visualization of HiFormer-B using Grad-CAM [5].



(a) Ground Truth (b) HiFormer-B (c) P^l Before DLF (d) P^l After DLF (e) P^s Before DLF (f) P^s After DLF

Figure 2. Feature visualization of HiFormer-B using Grad-CAM [5].