Beyond Self-Attention: Deformable Large Kernel Attention for Medical Image Segmentation
Supplementary Material

Reza Azad [1] Leon Niggemeier [1] Michael Hüttemann [1] Amirhossein Kazerouni [2]
Ehsan Khodapanah Aghdam [3] Yury Velichko [4] Ulas Bagci [4] Dorit Merhof [5]
[1]RWTH Aachen University [2]Iran University of Science and Technology
[3]Shahid Beheshti University [4]Northwestern University [5]University of Regensburg
{reza.azad, Leon.niggemeier, michael.huettemann@}@rwth-aachen.de, {dorit.merhof}@ur.de
{amirhossein477, ehsan.khpaghdam}@gmail.com, {ulas.bagci, y-velichko}@northwestern.edu

Abstract

The additional materials provided encompass an extended ablation study, which serves to demonstrate the robustness and efficacy of our method in addressing semantic segmentation tasks. Furthermore, we present supplementary visualizations and discussion that accentuate the impactful role played by the D-LKA module that we have proposed.

1. Computational complexity of the D-LKA

A comparison of the number of parameters for normal convolution and the constructed convolution is shown in table 1. While the numbers of the standard convolution explode for a larger number of channels, the parameters for decomposed convolution are lower in general and do not increase as fast. Deformable decomposed convolution adds a lot of parameters in comparison to decomposed convolution but is still significantly smaller than standard convolution. The main amount of parameters for deformable convolution is created by the offset network. Here, we assumed a kernel size of (5,5) for the deformable depth-wise convolution and (7,7) for the deformable depth-wise dilated convolution. This results in the optimal number of parameters for a large kernel of size 21×21. A more efficient way to generate the offsets would greatly reduce the number of parameters.

It is worth noting that the introduction of the deformable LKA does indeed introduce additional parameters and floating-point operations per second (FLOPS) to the model. However, it's important to emphasize that this increase in computational load does not impact the overall inference speed of our model. Instead, for batch sizes > 1, we even observe a reduction in inference time, shown in Figure 1. For instance, based on our extensive experiments, we have observed that for a batch size of 16, the inference times with and without deformable convolution are only 8.01ms and 17.38ms, respectively. We argue that this is due to the efficient implementation of the deformable convolution in 2D. To measure the times, a random input of size $(b \times 3 \times 224 \times 224)$ is used. The network is inferred 1000 times after a GPU warm-up period of 50 iterations. The measurements are done on an NVIDIA RTX 3090 GPU.

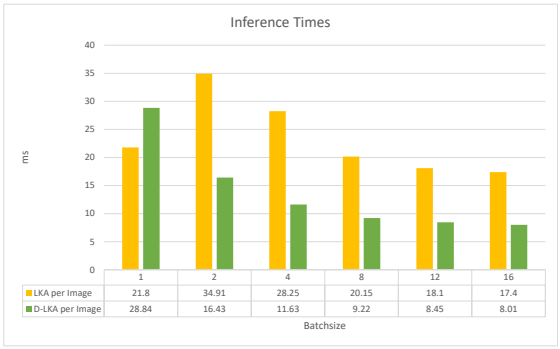

Figure 1. The inferences time in ms for an input image of size $3 \times 224 \times 224$ on the 2D methods. The times are already calculated for a single image for better comparison.

2. Performance vs Efficiency

To leverage the performance vs parameter tradeoff we visualize the performances on Synapse 2D dataset, reported in DSC and HD, and the memory consumption based on

Table 1. The number of parameters for standard convolution and decomposed convolution. The kernel size is 21×21. Adapted from [4].

# Channels	Std. Conv.	Decomp. Conv.	Deform. Decom. Conv.	Offset DDW-Conv.	Offset DDW-D Conv.
$C = 32$	$451,584$	$3,392$	$197,204$	$40,050$	$153,762$
$C = 64$	$1,806336$	$8,832$	$396,308$	$80,050$	$307,426$
$C = 128$	$7,225344$	$25,856$	$800,660$	$160,050$	$614,754$
$C = 256$	$28,901,376$	$84,480$	$1,633,940$	$320,050$	$1,229,410$
$C = 512$	$115,605,504$	$300,032$	$3,398,804$	$640,050$	$2,458,722$

the number of parameters in Figure 2. The D-LKA Net

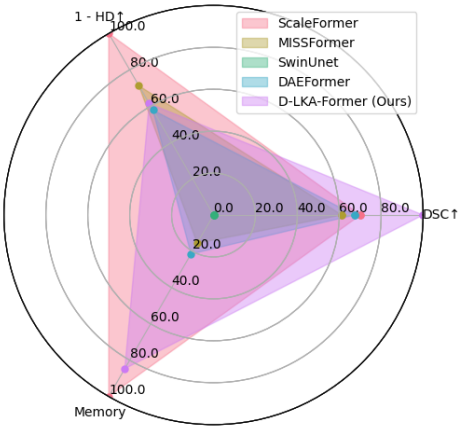

Figure 2. Performance vs memory chart to compare the performance of SOTA approaches, including ScaleFormer [5], MISSFormer [6], SwinUnet [2], DAEFormer [1], with our the proposed 2D D-LKA Net on Synapse dataset. DSC, HD, and Memory values are normalized using min-max normalization for improved visibility and comparability.

induces a rather large amount of parameters with approximately 101M. This is less than the second best-performing method, the ScaleFormer [5], which used 111.6M parameters. Compared to the more light-weight DAEFormer [1] model, we, however, achieve a better performance justifying the parameter increase. The majority of the parameters are from the MaxViT encoder; thus, replacing the encoder with a more efficient one can reduce the model parameters. It's also worth noting that in this visualization, we initially normalized both the HD and memory values within the [0, 100] range. Subsequently, we scaled them down from 100 to enhance the representation of higher values.

3. Qualitative results on the Synapse dataset

To further visualize our model capability, we provide a different perspective of the 3D organ segmentation of the Synapse dataset in Figure 3 and Figure 4. We neglect the visualization of the liver and the stomach so partly occluded organs get a better visibility.

To gain a better understanding of the limitations associated with the 2D approach, it is advisable to expand our perspective into the 3D domain. As illustrated in Figure 5, we can observe inconsistencies among the slices. These discrepancies can be attributed to the absence of information exchange between neighboring slices in a 2D network. In contrast, our 3D network successfully mitigates these limitations.

4. Limitations on the Skin dataset

Figure 6 shows a qualitative visualization of ISIC 2018 samples, where our approach fails. However, it is also visible that the segmentation is either noisy or quite primitive. Since this is also present in the training data, this could hinder the network from learning accurate segmentations.

5. Robustness Visualization

In line with the ablation study presented in the main paper, we conducted a thorough evaluation of various methods on the Synapse 2D dataset. To ensure the robustness of our findings, we executed each model five times and reported their statistical significance. This detailed analysis is visually represented in Figure 7.

References

[1] Reza Azad, René Arimond, Ehsan Khodapanah Aghdam, Amirhosein Kazerouni, and Dorit Merhof. Dae-former: Dual attention-guided efficient transformer for medical image segmentation. *arXiv preprint arXiv:2212.13504*, 2022. 2, 5

[2] 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, 2021. 2, 5

[3] 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. 5

[4] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022. 2

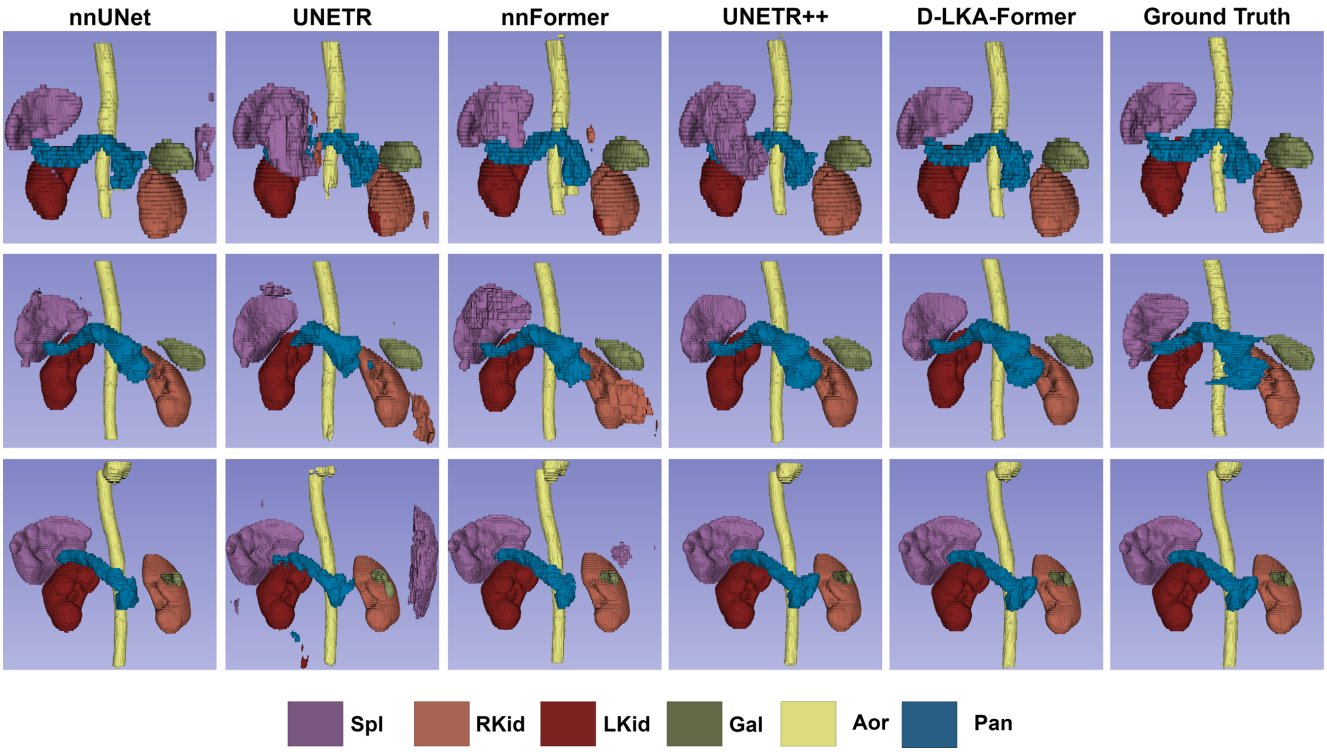

Figure 3. Additional qualitative results on the Synapse Dataset. Liver and Stomach are not shown for improved visibility of smaller occluded organs.

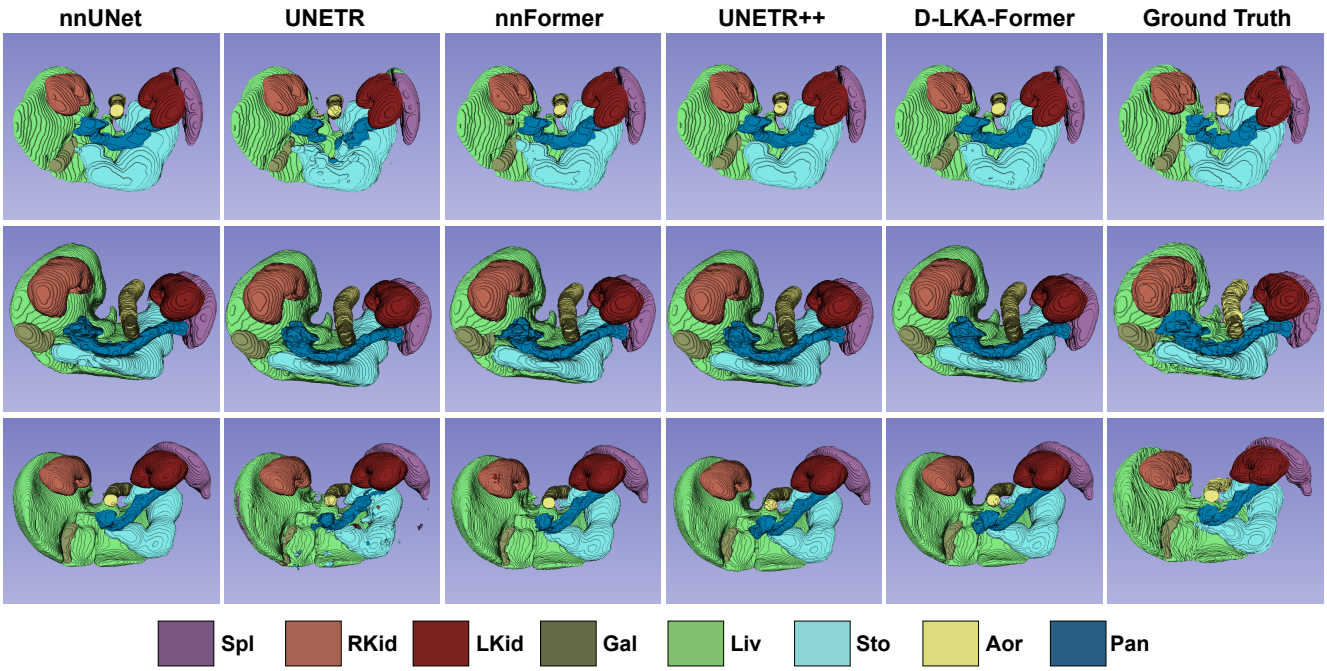

Figure 4. Additional qualitative results on the Synapse Dataset.

[5] Huimin Huang, Shiao Xie, Lanfen Lin, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Ruofeng Tong. Scaleformer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation. *arXiv preprint*

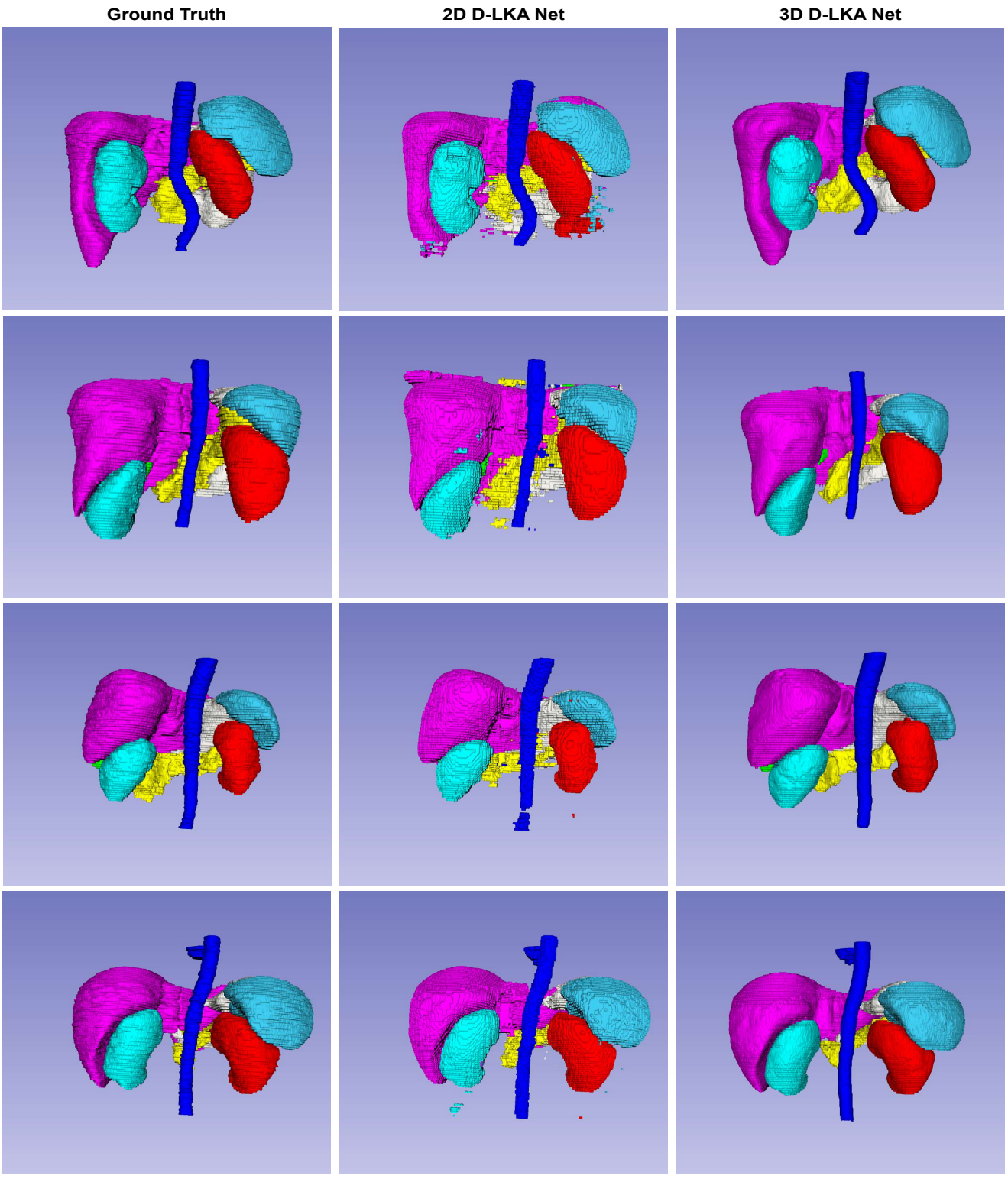

Figure 5. Additional qualitative results of the 2D D-LKA-Former on the Synapse dataset, visualized in 3D. The comparison to the 3D D-LKA Net is shown. Here, it is visible that the 3D version creates less noise due to the inter-slice dependencies.

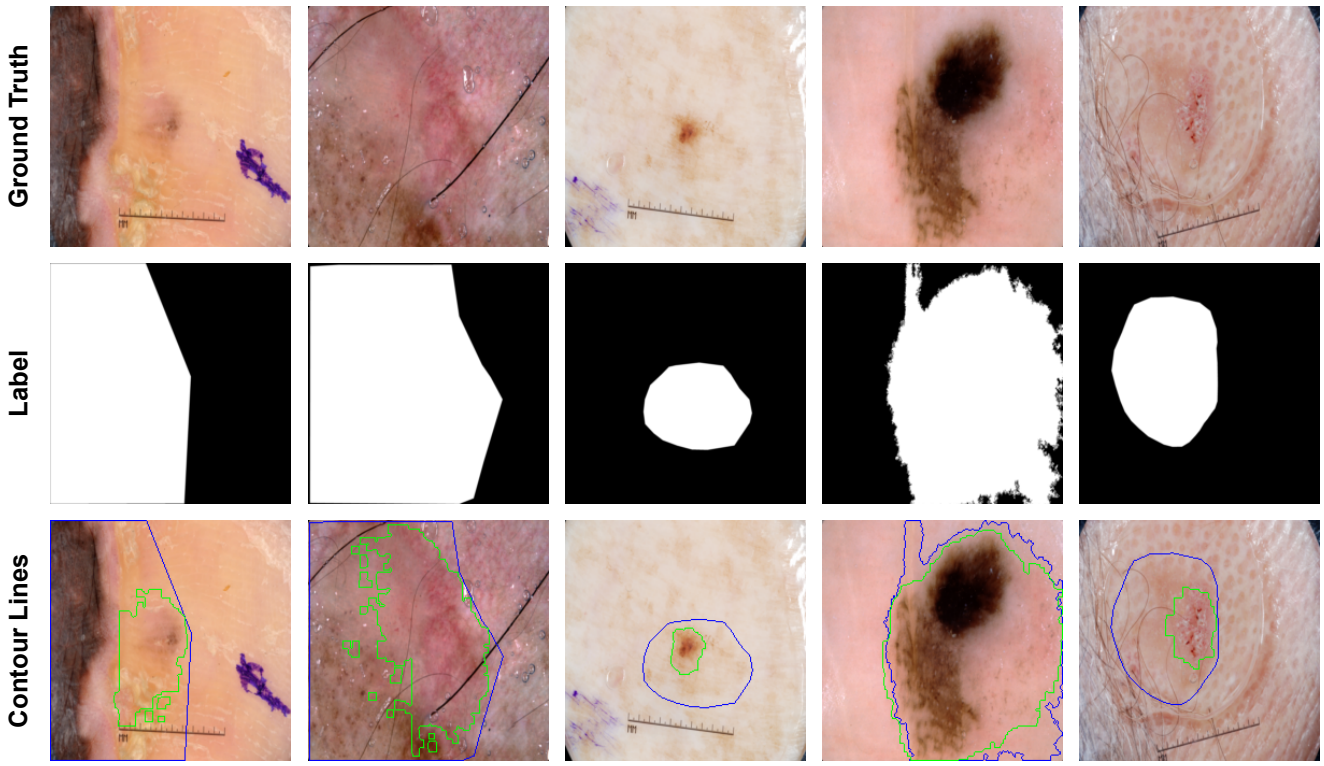

Figure 6. Additional qualitative results of the 2D D-LKA-Former on the ISIC 2018 dataset.

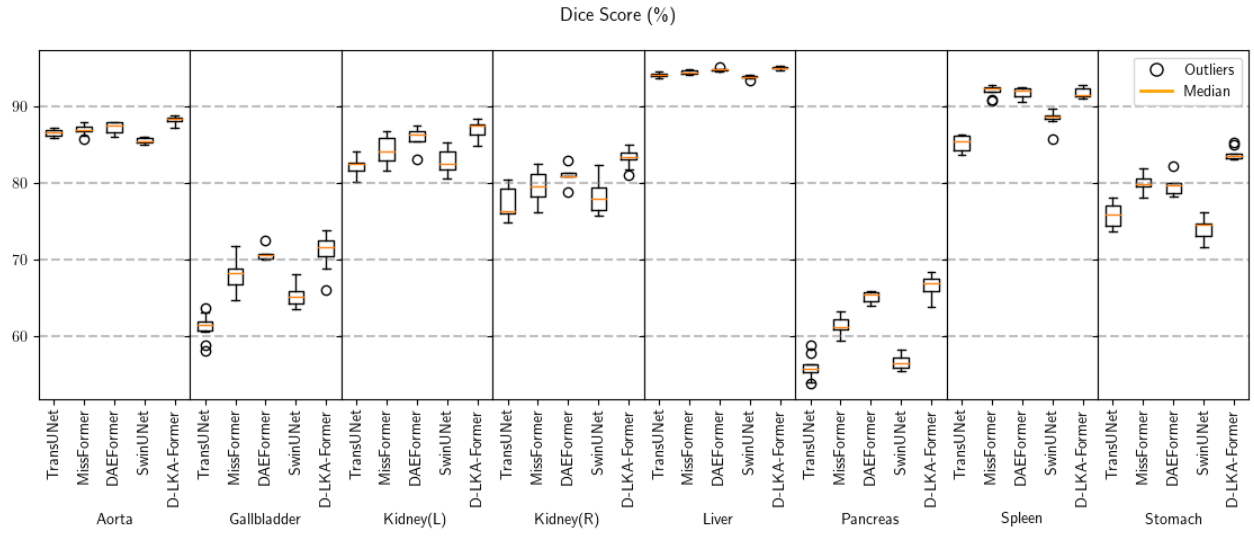

Figure 7. Statistical evaluation of single organ performance of Synapse dataset comparing state-of-the-art methods, including TransUnet [3], MISSFormer [6], SwinUnet [2], DAEFormer [1], with our the proposed 2D D-LKA Net. Visualized are the results on the Synapse dataset with the performance reported in DSC.

arXiv:2207.14552, 2022. 2

[6] Xiaohong Huang, Zhifang Deng, Dandan Li, Xueguang Yuan, and Ying Fu. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2022. 2, 5