DAST: Differentiable Architecture Search with Transformer for 3D Medical Image Segmentation

Dong Yang(✉), Ziyue Xu, Yufan He, Vishwesh Nath, Wenqi Li,
Andriy Myronenko, Ali Hatamizadeh, Can Zhao, Holger R. Roth,
and Daguang Xu

NVIDIA, Santa Clara, USA
dongy@nvidia.com

Abstract. Neural Architecture Search (NAS) has been widely used for medical image segmentation by improving both model performance and computational efficiency. Recently, the Visual Transformer (ViT) model has achieved significant success in computer vision tasks. Leveraging these two innovations, we propose a novel NAS algorithm, DAST, to optimize neural network models with transformers for 3D medical image segmentation. The proposed algorithm is able to search the global structure and local operations of the architecture with a GPU memory consumption constraint. The resulting architectures reveal an effective relationship between convolution and transformer layers in segmentation models. Moreover, we validate the proposed algorithm on large-scale medical image segmentation data sets, showing its superior performance over the baselines. The model achieves state-of-the-art performance in the public challenge of kidney CT segmentation (KiTS'19).

Keywords: Neural architecture search · Transformer · Segmentation

1 Introduction

Image segmentation is one of the most fundamental and popular tasks in medical image analysis. It is widely applied to parse organs, bones, soft tissues, or lesions in *N*-*D* medical images. Conventional methods rely on statistics of image intensities or object shapes to infer the boundaries of the target regions [8]. Recently, the convolutional neural networks (CNN) demonstrated superior performance in multiple tasks. CNNs are inherently translation-invariant with in-neighborhood computation, which makes training efficient and deployment effective. For instance, the U-shaped encoder-decoder CNN is greatly favored among segmentation models due to its simplicity and effectiveness [16]. Despite the success, adding new components to existing models in pursuit of better performance and efficiency is always an ongoing effort in different research fields.

Inspired by the advancement from related domains, e.g., natural language processing (NLP), transformers [24] have been successfully introduced to image

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 747–756, 2023.
https://doi.org/10.1007/978-3-031-43898-1_71

processing and computer vision [7]. The transformer block is crafted with long-range dependency inside sequences with marginal inductive bias. To incorporate a transformer into image analysis models, images are divided into patches with equal size and serialized as a sequence of tokens, so that the transformer based models can treat N-D images in the same way as 1-D sentences. Such operation explicitly destroys neighborhood relationships in the images, and instead learnable position embeddings are added to encourage the learning of flexible patch interaction. In addition to supervised tasks, transformer-based models have also been shown to achieve superior performance in pre-training with large-scale (labeled/unlabeled) data sets [11].

Most existing neural architectures are designed with strong human heuristics. Neural architecture search (NAS) has been proposed in an attempt to reduce dependency on such heuristics while optimizing model performance for given tasks. Given target constraints, it is capable of optimizing multiple objectives (e.g., accuracy, memory consumption, latency, etc.) of the neural network models at the same time. Nowadays, NAS has been widely applied for many applications in medical imaging including image classification and segmentation.

Existing NAS works have been focusing on optimization with convolutional deep learning components [18,31]. Our proposed NAS method, named DAST, on the other hand learns the relationship between convolutions and transformers within the search space of segmentation networks. During architecture searching, those two operations can be placed at different scale resolutions and levels for performance optimization. Intrinsically, it shall benefit from inductive biases of these two popular deep learning ingredients. Meanwhile, DAST is also equipped with capacities of optimizing memory consumption of the searched architecture, so that the input shape of the neural network can be properly adjusted according to the available computing resource, and long-range dependency of transformers can be visualized through attention matrices. We evaluated our proposed algorithm on two public data sets with excellent performance.

2 Related Work

Neural architecture search tries to find optimal global model structures and local operations from large search spaces for different applications. Searching algorithms, including reinforcement learning and genetic algorithms [2,27,31], have been proposed for different search spaces. These approaches usually require large-scale computing resources to train a large number of independent neural networks, which makes them less practical when applied to large-scale data sets. On the contrary, differentiable neural architecture search (DARTS) aims to boost search efficiency and reduce computation budgets via continuous relaxation in the optimization [6,9,12,17,18,26,30]. It defines a large super-net containing all network candidates with learnable intermediate path weights, such that optimizing model architecture is equivalent to optimizing and binarizing those path weights. However, most existing NAS algorithms in medical imaging rely heavily on fully convolution-based search spaces, which may limit its receptive field.

Transformer based neural network has been recently introduced to medical imaging domain for various applications following the success of vision transformer (ViT) in computer vision [7]. The direct extension applies ViT to 2D medical image analysis [21,23]. Some works have adopted ViT for 3D medical image segmentation [3,4,10,22,25,29] via serializing 3D images as sequences of patches/cubes. Most works rely on conventional designs of neural architectures and replace the convolution operations with transformers. For instance, the segmentation networks are always in "symmetric" encoder-decoder structure. However, from a network design's perspective, it is not trivial to find the right balance between convolutions and transformers inside the architecture.

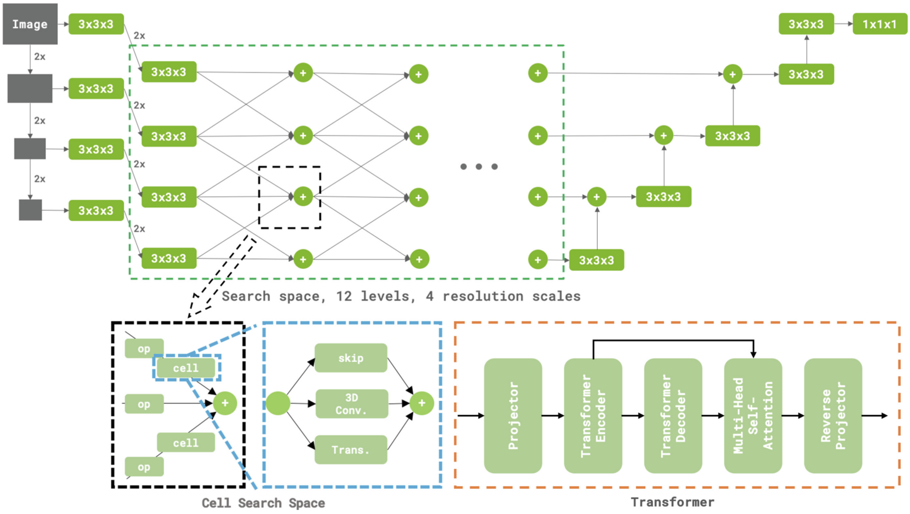

Fig. 1. Search space of global structure and local operations in DAST. The bottom right figure shows the transformer layer in our cell search space.

3 Method

Our NAS algorithm, DAST, is an intuitive extension of DiNTS [12] for 3D medical image segmentation. Like other DARTS type of algorithms, it requires continuous relaxation of the super-net search space for gradient descent optimization. Unlike other NAS algorithms, it searches for a *multi-path* neural network and optimizes the searched architecture with additional memory constraints. Large image inputs can fit the searched architecture with low memory consumption, which helps to build long-range dependencies for transformers.

Differentiable Search Space. Following DiNTS [12], the main search space is defined as $R \times L$ grids with R resolution scales and L levels from input image to output segmentation shown in Fig. 1. The grid node $n_{i,j}$ (at resolution i and level j) has directed connections towards neighboring nodes $n_{i-1,j+1}$, $n_{i,j+1}$

and $n_{i+1,j+1}$. Each edge e of connection is a weighted combination of outputs from operations o, and the pool of o includes skip-connection, convolution and transformer. For neighboring nodes at different levels, additional ×2 up-sampling or down-sampling is added in e, as well as convolutions to match feature channel numbers. There is no additional operation when passing features between nodes at the same level. To optimize the architecture, two different types of weights are introduced. Each edge has weight w_e, and each operation has weight w_o. Then the searched architecture can be defined when all w_e and w_o are binarized.

The stem cells are concatenated at input and output of the search space. The input stem cells down-samples image toward different resolution scales and fit image features to the search space. The output stem cells up-samples multi-scale features out of search space with necessary concatenation to produce multi-channel probability maps.

Transformer. We introduce transformer [24] into the search space as a candidate of the operation pool as shown in Fig. 1. The input and output of transformer are with dimension $C \times N$ (C is feature dimension and N is length of the sequence). However, this dimension is not suitable for networks with high dimensional image features. Therefore, we add a convolutional projector $\mathcal{P}$ [5] before the transformer, aiming to resize and project features $\mathcal{X}^{c \times h \times w \times d}$ to a smaller size with the number of channels matching the number of tokens required by the transformer (h, w, d denotes height, width, depth of 3D patches). Another input of projector is a learnable 3D positional encoding P shown in Eq. 1.

$$\mathcal{X}_{\text{in}} = \mathcal{P}\left(\text{Concat}\left(\mathcal{X}, P\right)\right) \tag{1}$$

The positional embedding P in Eq. 2 is initialized as a normalized 3D position map. It is efficient to compute with only 3 channels.

$$\begin{aligned} P\left[0,i,j,k\right] &= i/h, \quad i \in [0, h-1] \\ P\left[1,i,j,k\right] &= j/w, \quad j \in [0, w-1] \\ P\left[2,i,j,k\right] &= k/d, \quad k \in [0, d-1] \end{aligned} \tag{2}$$

The full transformer with an encoder and a decoder is adopted to process the output of the projector. The decoder takes additional learnable query embedding as input. The output of the transformer shares the same dimension/shape with the input. Next, we need to add another reversed projector to map the 1-D feature map back to the 3D shape of input.

Segmentation Attention. To further understand the self-attention scheme, we embed an additional multi-head self-attention layer $\mathcal{A}$ after the transformer. $\mathcal{A}$ uses the feature maps from the transformer as the semantic query q, and the features from the transformer encoder as key k. Unlike multi-head self-attention layers inside the transformer, $\mathcal{A}$ **does not have residual connection**. Thus, $\mathcal{A}$ is enforced to learn meaningful attention weights for segmentation tasks. The attention weights are directly multiplied with intermediate features maps, which can be reshaped from 1-D to 3-D for visual interpretation.

Memory Estimation. Like DiNTS [12], the memory budget constraints were proposed as part of loss functions to optimize memory usage at training and

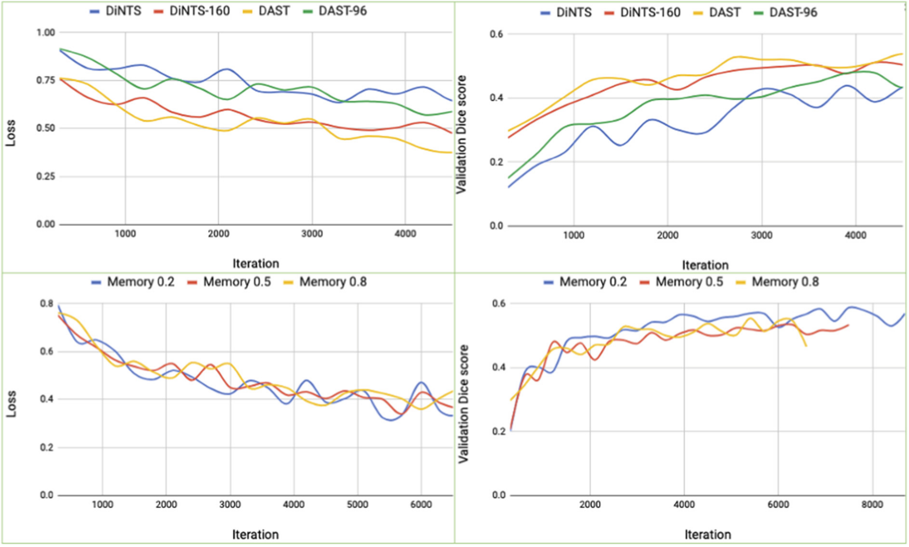

Fig. 2. Top: training and validation curves of DiNTS and DAST architectures when re-training; bottom: training and validation curves of DAST architectures when re-training with different memory constraints (λ).

inference. It requires to estimate peak memory usage in each operation given the fixed input shape. Since the token is designed to have the same dimension C as the channel dimension at different resolution scales, the memory consumption estimation of entire transformer is shown in Eq. 3:

$$M_{\text{transformer}} = N \times C \times (l')^3 \tag{3}$$

In practice, the number of token l' is fixed as 512 to avoid potential memory explosion. Eight heads are used in each multi-head self-attention layer. Number of operations N is approximately estimated as 15 including convolutions, batch normalization, linear operations, layer normalization, and multi-head self-attention.

4 Experiments

4.1 Data Sets and Implementation Details

We adopted large-scale data sets Task07 Pancreas from Medical Segmentation Decathlon (MSD) [1] as used in [12] for architecture searching, and KiTS'19 [13,14] to validate the searched architectures. Both are very challenging applications involving various fields of view of CT volumes and different types of pathology. The pancreas data set has 3-class segmentation labels (background, pancreas and tumor) for 282 CT volumes. We adopt entire labeled set for NAS with the same data split as [12]: 114 volumes for model training, 114 volumes for architecture search, and 54 volumes for model validation. A different $4:1$ data split is used for experiments of training from scratch. The KiTS'19 data set has 3-class segmentation labels (background, kidney and tumor) for 210 CT volumes, and an additional standalone 90 test volumes (with hidden ground truth) for the public leaderboard. We train the searched models with 5-fold data split, and verify the model performance on the test set using the public leaderboard. All data sets are re-sampled into the isotropic voxel spacing 1.0 mm for both images and labels. For both CT data sets, the voxel intensities of the images are normalized to the range $[0, 1]$ according to the 5^{th} and 95^{th} percentile of overall foreground intensities.

A combination of dice loss and cross-entropy loss is adopted to minimize both global and pixel-wise distance between ground truth and predictions. The NAS is conducted through conventional bi-level optimization following implementation[1]. We use the same definition of search space with 4 resolution scales and 12 levels. For cell level searching, we only use three different operations: skip-connection, convolution and transformer. For model training, we use a large input patch shape 160^3, and the batch size at each GPU is 2. The training settings (like data augmentation, optimizer, etc.) are very similar to the model searching. We keep a constant learning rate $1e^{-3}$ to train the model from scratch for $40,000$ iterations. Our experiments are conducted using MONAI and trained on eight NVIDIA V100 GPUs with 32 GB of memory. For searching or training, the time cost is $\sim$ 15 hours including training and validation on-the-fly (similar as [12]).

4.2 Comparison with DiNTS

Since DiNTS is the closest work to DAST, we directly compare the performance on DiNTS's searching tasks with the same data split. Then we re-train the searched architectures from both methods from scratch. As shown in Fig. 2, DAST has better training convergence and validation accuracy compared to DiNTS. The default model input shapes of DAST and DiNTS are different, so we experiment with various combinations. With input shape 160^3, DAST converges faster than DiNS-160 with a better validation curve. The same conclusion can be made with input shape 96^3. Based on the results, training with smaller input shape would make the training process harder. DAST consistently has better performance than DiNTS under different settings.

4.3 KiTS'19 Experiments

To verify the effectiveness and generalization of our searched architectures from DAST, we validate the searched architecture (from pancreas data set) on this challenging task. Metrics for kidneys and tumors are the average Dice score per case. Finally, we evaluate our single-fold model as well as the ensemble from 5 cross-validation models on the public test leaderboard[2].

[1] https://github.com/Project-MONAI/tutorials/tree/master/automl/DiNTS.

[2] https://kits19.grand-challenge.org/evaluation/challenge/leaderboard/.

Table 1. KiTS'19 challenge test-set performance evaluation for kidney and tumor segmentation in terms of the average Dice score per case. The evaluation results of our method are copied directly from the public leaderboard.

Method	Kidney Dice	Tumor Dice	Average
3D U-Net [19]	0.9730	0.8250	0.8990
Cascaded 3D U-Net [28]	0.9740	0.8310	0.9025
VB-Net [20]	0.9730	0.8320	0.9025
Cascaded 3D U-Net [15]	0.9670	0.8450	0.9060
nnU-Net (20 U-Net models & ensemble) [13]	0.9740	0.8510	0.9125
nnU-Net (20 U-Net models & ensemble) [16]	-	0.8542	-
DAST (1 model)	**0.9774**	0.8522	**0.9148**
DAST (5 models & ensemble)	**0.9799**	**0.8568**	**0.9184**

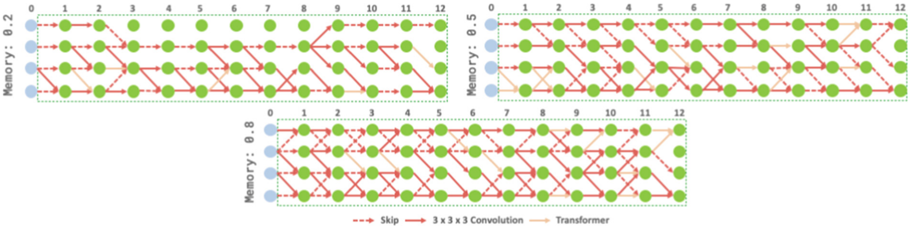

Fig. 3. Searched architectures under different memory constraints ($\lambda = 0.2, 0.5, 0.8$).

Based on the results from the public leaderboard in Table. 1, our single-fold model and ensemble of five models achieve excellent performance compared to all other entries in the challenge shown in the Table. 1. The nnU-Net [16] is the best among all other entries, but the method utilized 20 U-Net models with training strategies to achieve the ensemble result for the challenge. Some other entries rely on cascaded models, which use more complex and intensive training mechanisms. On the contrary, DAST shows great simplicity when transferring a searched architecture to a new task. It is important to point out that the performance of our models is not only the best of all entries with publications, but also the best of all public entries (around $2,000$) on the test leaderboard.

4.4 Ablation Studies

Memory Constraints. We provide the option to change the parameter controlling memory consumption budget in the loss function with different values shown in Fig. 3. From the results, we can observe that given different values, the searched models have a clear trend: for the model with the highest memory ($\lambda = 0.8$), transformers are distributed at different resolution scales. As memory constraints increase (where λ is reduced), transformers are more towards lower

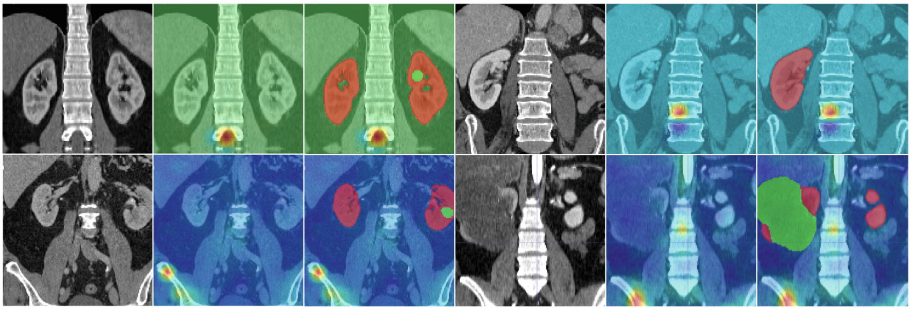

Fig. 4. Four attention weights visualized with CT images and model predictions. The first row is from a transformer layer ($r = 4$, $l = 8$), and the second row is from another transformer layer at higher resolution ($r = 3$, $l = 10$). In each case, the left side is the original image, the middle one is the overlaid display with the attention weights, and the right side is the overlaid display of the attention weights and segmentation masks.

resolution scales. It agrees with our expectation since transformers normally consume more GPU memory than convolutions due to several linear operations in large token dimension. On the other hand, more convolutions are chosen than transformers, which implicitly suggests that this balance between convolutions and transformers is better for feature learning in segmentation. Another benefit shown in Fig. 2 is that re-training architectures with lower memory constraints would not hurt the model performance. It is encouraging to see that such combination of convolution and transformer retains the same-level performance with lower GPU memory costs and receptive field of the entire model input.

Attention Mechanism. We visualize the attention weights computed by a dedicated self-attention operation in the transformer. The attention weight is with shape $512 \times 4096 = 512 \times 16^3$. Then we take the average from the channel dimension and resize it to a volume with shape 160^3 by trilinear interpolation (for visualization). We can see that the self-attention weights of the kidney segmentation consistently focus on the lower spine or pelvis areas at different transformer layers as evidence of long-range dependency (Fig. 4). One potential explanation could be that kidneys are located around those areas and both kidneys are on the opposite sides of the spine. So the information over there can help roughly identify the kidneys from the whole-body CT. Especially specific bones are good bio-markers with high intensity values in CT. Furthermore, to the best of our knowledge, it is the first time that the multi-head self-attention is visualized for 3D medical image segmentation.

5 Discussion and Conclusion

In this study, we observe that DAST is able to find effective and concise relationships between convolutions and transformers in a single neural network model. The optimized connections between operations improves the model effectiveness in various applications. Such models benefit from the different inductive biases introduced by these two operations. Adding a memory constraint loss as an additional objective can lower memory consumption for the searched architecture. Transformers will then benefit more from long-range dependencies of larger input patches. We hope this perspective will be helpful for different applications in medical imaging.

References

1. Antonelli, M., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
2. Bae, W., Lee, S., Lee, Y., Park, B., Chung, M., Jung, K.-H.: Resource Optimized Neural Architecture Search for 3D Medical Image Segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11765, pp. 228–236. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32245-8_26
3. Cao, H., et al.: Swin-unet: unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
4. Chen, J., et al.: TransuNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Ding, M., et al.: HR-NAS: searching efficient high-resolution neural architectures with lightweight transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2982–2992 (2021)
6. Dong, N., Xu, M., Liang, X., Jiang, Y., Dai, W., Xing, E.: Neural Architecture Search for Adversarial Medical Image Segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 828–836. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_92
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Elnakib, A., Gimel'farb, G., Suri, J.S., El-Baz, A.: Medical image segmentation: a brief survey. Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies pp. 1–39 (2011)
9. Guo, D., et al.: Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4223–4232 (2020)
10. Hatamizadeh, A., et al.: UNETR: transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584 (2022)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021)
12. He, Y., Yang, D., Roth, H., Zhao, C., Xu, D.: Dints: differentiable neural network topology search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5841–5850 (2021)
13. Heller, N., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: results of the kits19 challenge. Med. Image Anal. **67**, 101821 (2020)
14. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)

15. Hou, X., Xie, C., Li, F., Nan, Y.: Cascaded semantic segmentation for kidney and tumor. Submissions to the (2019)
16. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
17. Kim, S., et al.: Scalable Neural Architecture Search for 3D Medical Image Segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 220–228. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_25
18. Liu, H., Simonyan, K., Yang, Y.: Darts: differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
19. Ma, J.: Solution to the kidney tumor segmentation challenge 2019 (2019)
20. Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y.: Segmentation of kidney tumor by multi-resolution VB-Nets (2019)
21. Park, S., Kim, G., Kim, J., Kim, B., Ye, J.C.: Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis. Adv. Neural Inf. Process. Syst. **34** (2021)
22. Tang, Y., et al.: Self-supervised pre-training of swin transformers for 3d medical image analysis. arXiv preprint arXiv:2111.14791 (2021)
23. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 36–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_4
24. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
25. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 171–180. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_16
26. Yan, X., Jiang, W., Shi, Y., Zhuo, C.: MS-NAS: Multi-scale Neural Architecture Search for Medical Image Segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12261, pp. 388–397. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_38
27. Yu, Q., et al.: C2FNAS: coarse-to-fine neural architecture search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4126–4135 (2020)
28. Zhang, Y., et al.: Cascaded volumetric convolutional network for kidney tumor segmentation from CT volumes. arXiv preprint arXiv:1910.02235 (2019)
29. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnFormer: interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
30. Zhu, Z., Liu, C., Yang, D., Yuille, A., Xu, D.: V-NAS: neural architecture search for volumetric medical image segmentation. In: 2019 International conference on 3d vision (3DV). pp. 240–248. IEEE (2019)
31. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578 (2016)