5-5-2022

Segmentation with Super Images: A New 2D Perspective on 3D Medical Image Analysis

Ikboljon Sobirov

Numan Saeed

Mohammad Yaqub

Segmentation with Super Images: A New 2D Perspective on 3D Medical Image Analysis

Ikboljon Sobirov, Numan Saeed, and Mohammad Yaqub

Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{ikboljon.sobirov, numan.saeed, mohammad.yaqub}@mbzuai.ac.ae

Abstract. Deep learning is showing an increasing number of audience in medical imaging research. In the segmentation task of medical images, we oftentimes rely on volumetric data, and thus require the use of 3D architectures which are praised for their ability to capture more features from the depth dimension. Yet, these architectures are generally more ineffective in time and compute compared to their 2D counterpart on account of 3D convolutions, max pooling, up-convolutions, and other operations used in these networks. Moreover, there are limited to no 3D pretrained model weights, and pretraining is generally challenging. To alleviate these issues, we propose to cast volumetric data to 2D super images and use 2D networks for the segmentation task. The method processes the 3D image by stitching slices side-by-side to generate a super resolution image. While the depth information is lost, we expect that deep neural networks can still capture and learn these features. Our goal in this work is to introduce a new perspective when dealing with volumetric data, and test our hypothesis using vanilla networks. We hope that this approach, while achieving close enough results to 3D networks using only 2D counterparts, can attract more related research in the future, especially in medical image analysis since volumetric data is comparably limited.

Keywords: Super Images · 2D Analysis · 3D Analysis · Biomedical Volumetric Image Segmentation

1 Introduction

The latest strides in deep learning (DL) and convolutional neural networks (CNNs) in particular, have proven its viability in numerous computer vision tasks such as classification [16], detection [15,3] and segmentation [18,11]. Its reach in medical applications as well as the availability of datasets for a multitude of medical cases paved the way for the development and increased use of CNNs in the field. There exists a huge potential to using DL in medical imaging as it can not only automate this process, speeding up clinicians' work, but also yield results comparable to those of doctors. This gives a chance to clinicians to redirect their focus on more critical tasks.

Multiple DL solutions have been proposed for each of the medical tasks. In medical image segmentation, U-Net [22] is considered the go-to CNN in this

field. Later models following that have introduced certain variations [6,13,14,31] to U-Net until transformers [27] came along. Vision transformer (ViT) [10] was a game-changing architectural design initially designed for classification, and not far until applied in segmentation tasks [5,26,30].

Three schools of thought have emerged on account of the nature of medical images, which is different from natural images. The distinguishing nature here refers to the three-dimensionality attribute of many medical images, such as computed tomography (CT) [7], positron emission tomography (PET) [9], magnetic resonance imaging (MRI) [19], and so forth.

The first group of researchers [1,6,20] support using 3D datasets primarily for the fact that 3D networks are able to capture depth information that 2D counterpart lacks, claiming that this information is crucial to model learning. Another argument is that the nature of 3D data is closer to that of real life, which is why 3D models perform better [1,20]. On the downside, instead of 2D, 3D convolutions, max pooling, up-convolutions are applied during model learning, thus requiring much more compute power and training/inference time [4,12].

On the other hand, those who argue that 2D should still be in heavy use reinforce their claim that utilizing 2D images is not only more cost- and time-effective, but also offers more options to apply transfer learning [12,23,29]. Transfer learning, with weights pretrained on large-scale datasets, such as ImageNet [8], can be considerably beneficial to model learning especially when datasets are small in quantity. Another advantage of using 2D images is the fact that 3D can be easily converted to multiple 2D slices, generating a larger scale set compared to relying on a limited number of 3D counterparts. Another vivid upside is that there are numerous 2D architectures available for the encoder of U-Net [6] like models [21]. This makes 2D models easier to customize and adjust to the need of the problem at hand.

The third party [17,23,25] represents the hybrid approach of 2D and 3D in various fashions. Seemingly, this approach should combine the best of 2D and 3D models, yet in reality, this is not always the case, fundamentally failing to capture innate benefits of 2D and 3D architectures [29].

In this paper, we introduce a new perspective to using volumetric data in a 2D fashion. We generate 2D super images (SIs) from 3D input by stacking the depth information (i.e., slices) side-by-side, and training a 2D model for the same task. A similar work was proposed by [12] on natural videos; unlike them, this concept is newly introduced to the medical field, and to volumetric data not video frames. This novel approach in the segmentation task, achieving comparable results to a 3D model counterpart, can lay the foundation for a new look into dealing with 3D medical data. As contributions of our paper:

- We introduce a new perspective to dealing with biomedical volumetric data by casting them into super images and training them with 2D models;
- We empirically show that pretraining and preprocessing techniques can easily boost the performance of the models that use super images;

– We validate our approach using datasets of CT, PET, and MRI and different tissue types as tumor and atria so as to show the effectiveness of the approach.

2 Methodology

2.1 Datasets & Preprocessing

Head and Neck Tumor: To validate our new approach, we experimented with two different datasets: head and neck tumor segmentation and outcome prediction (HECKTOR) challenge [2] and atrial segmentation challenge [28] datasets. HECKTOR dataset comprises 224 CT and PET scans of patients with head and neck tumor for the training set (i.e. dataset is available online[1]). Bounding box information comes with the dataset for localization of the tumor region, which was used to crop the scans and the mask down to the size of $144 \times 144 \times 144 mm^3$ with consistency between the scans. Further cropping down to $80 \times 80 \times 48$ was performed around the tumor region for faster and better performance as in [24]. This highlights the tumor region, allowing the models to learn more easily. Since the bounding box information is provided by the challenge organizers, tumor is within the cropped region, and mappings between both modalities and mask are accurate. Further preprocessing techniques applied were re-sampling the data to have isotropic voxel spacing, and the intensity normalization of both CT and PET data.

Atria: Atrial segmentation challenge dataset includes 100 3D gadolinium contrast (GE) MRIs for the training set[2]. The scans are of different dimensions, thus were resized to the same size of $512 \times 512 \times 88 mm^3$. Similarly, intensity normalization was applied on the scans.

No further data augmentations were applied on either dataset, unless reported otherwise, for a minimalistic approach. The testing set ground truth is inaccessible in both datasets, therefore, they are not used, and instead, k-fold cross validation was utilized for all the experiments.

We purposely chose two datasets of different modalities (i.e. CT and PET in one, and MRI in another) to test the hypothesis for generalizability. Moreover, the tasks are the segmentation of tumor in one dataset and atria in the other dataset. This shows that the idea is not limited to a specific tissue type with specific characteristics.

2.2 Super Image Generation

3D volume can provide features from the depth information for the model to learn since they use 3-dimensional kernels, but we expect that these characteristics are still detectable and learnable in 2D SIs by well-designed deep neural networks.

[1] aicrowd.com/challenges/miccai-2021-hecktor
[2] atriaseg2018.cardiacatlas.org/data

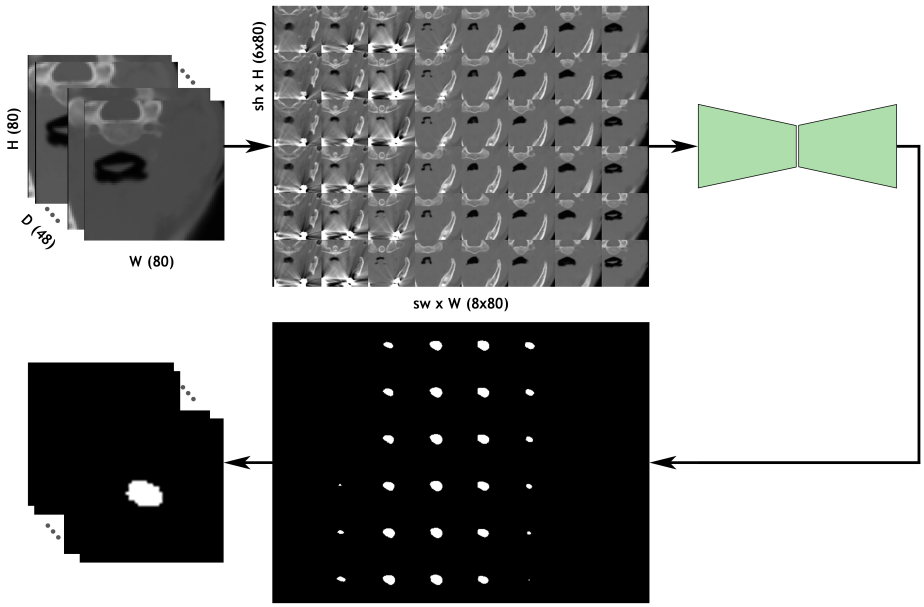


Fig. 1: The figure shows the construction of super images from volumetric data. We rearrange the depth dimension by assembling the slices together to generate the super image. It is then fed to a 2D segmentation network. The model yields the prediction mask which is then rearranged back to the original shape. Note that the volumetric prediction mask shows a tumor region for visualization purposes.

With that in mind, we generate SIs from volumetric data by taking slices and stitching them together side by side in order as shown in Figure 1. Given a 3D image $x \in \mathbb{R}^{H \times W \times D \times C}$, where H is the height, W is the width, D is the depth, and C is the number of channels, the depth dimension is rearranged. The resulting image $s \in \mathbb{R}^{\hat{H} \times \hat{W} \times C}$ is now 2D, where $\hat{H} = H \times sh$, and $\hat{W} = W \times sw$; sh and sw represent the degree by which the height and width should be rearranged respectively to generate a grid size of $sh \times sw$. As a demonstration, the size of $80 \times 80 \times 48 \times 2$ (2 for both CT and PET slices), having 48 as the depth, can be considered with sh of 6 and sw of 8, thus generating the SI in the dimensions of $480 \times 640 \times 2$, as shown in Figure 1. 2D U-Net (or any other 2D segmentation model for that matter) can then be applied on these SIs to perform the segmentation.

2.3 Implementation Details

For our experiments, we used two NVIDIA RTX A6000 (48GB) GPUs, and the implementation was done utilizing the PyTorch library. We ran all the experiments for 100 epochs. An AdamW optimizer with the initial learning rate of

Table 1: The table shows the results of vanilla 3D U-Net (comparison target) to SI-based 2D U-Net on the HECKTOR training/validation dataset. The results are the mean and standard deviation of 5-fold cross validation.

Model	Image Size	*sh*	*sw*	DSC	Precision	Recall
3D U-Net	$144 \times 144 \times 144$	-	-	0.718±0.055	0.749±0.050	0.747±0.0675
2D U-Net	$144 \times 144 \times 144$	12	12	0.700±0.070	0.731±0.046	0.731±0.075
3D U-Net	$80 \times 80 \times 48$	-	-	**0.779**±0.031	0.787±0.021	0.822±0.039
2D U-Net	$80 \times 80 \times 48$	8	6	**0.778**±0.033	0.799±0.021	0.810±0.044
2D U-Net	$80 \times 80 \times 48$	6	8	0.777±0.034	0.793±0.018	0.816±0.044
2D U-Net	$80 \times 80 \times 48$	12	4	0.770±0.030	0.809±0.037	0.801±0.055
2D U-Net	$80 \times 80 \times 48$	4	12	0.759±0.043	0.790±0.016	0.797±0.062
2D U-Net	$80 \times 80 \times 48$	24	2	0.744±0.044	0.765±0.027	0.809±0.047
2D U-Net	$80 \times 80 \times 48$	2	24	0.762±0.035	0.779±0.023	0.809±0.052

0.001, and weight decay of 0.00001 was used, and a cosine annealing schedule that starts with the initial learning rate, decreasing it to the base learning rate of 0.00001, and resetting it after every 25 epochs was chosen to control the learning rate. The batch size was set to 4, 8, and 16 depending on the architecture and dataset. The evaluation metric was primarily a dice similarity coefficient (DSC), and additional precision and recall were also calculated.

2.4 Experiments and Results

To verify this new approach to dealing with volumetric data, we used two different datasets. For the HECKTOR dataset, two settings were explored: (i) with the initial size of $144 \times 144 \times 144 mm^3$, and (ii) the cropped size of $80 \times 80 \times 48 mm^3$.

Table 2: The table shows the results of vanilla 3D U-Net (comparison target) to SI-based 2D U-Net on the atrial segmentation training/validation dataset. The results are the mean and standard deviation of 4-fold cross validation. PT stands for 2D U-Net pretrained on ImageNet1k, and A stands for augmentations.

Model	Image Size	*sh*	*sw*	DSC	Precision	Recall
3D U-Net	$512 \times 512 \times 88$	-	-	**0.893**±0.011	0.898±0.011	0.894±0.024
2D U-Net	$512 \times 512 \times 88$	11	8	0.812±0.047	0.902±0.038	0.785±0.050
2D U-Net	$512 \times 512 \times 64$	8	8	0.851±0.039	0.913±0.018	0.822±0.063
PT	$512 \times 512 \times 64$	8	8	**0.895**±0.013	0.872±0.092	0.878±0.035
PT&A	$512 \times 512 \times 64$	8	8	**0.901**±0.008	0.919±0.018	0.890±0.029

For the HECKTOR dataset, 5-fold cross validation was used, and the mean and standard deviations are reported in Table 1. In the first setting, both the 3D U-Net with volumetric data and 2D U-Net with SIs were training using random initialization. The 3D U-Net achieved the DSC score of 0.718, precision of 0.749,

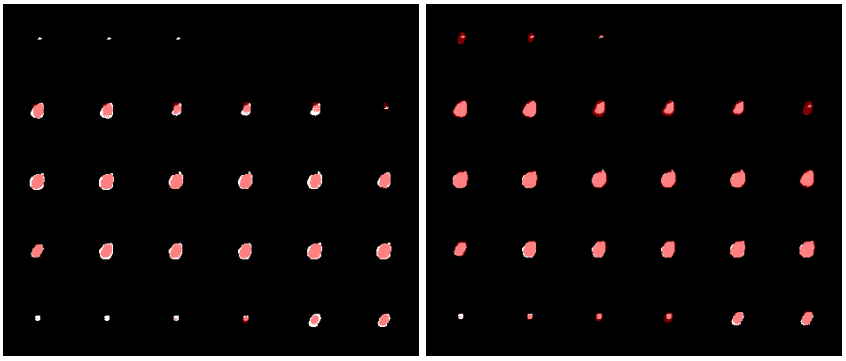
and recall of 0.747. The 2D SI-based network was able to reach the DSC of 0.700, which is only marginally lower than the 3D counterpart. The precision and recall were 0.731 and 0.731 respectively.

In the small size setting, a vanilla 3D U-Net, that was trained as the target source against the SI approach, reached the mean DSC of 0.779, precision of 0.787, and recall of 0.822. In the 2D setting, the values for sh and sw were set to different combinations to generate varied sized rectangular SIs as shown in Table 1. We rearranged the volumetric data with sh and sw set at 8×6, 6×8, 12×4, 4×12, 24×2, and 2×24 grid sizes, and trained 2D U-Net. Note that the selected combinations are the multiples of the depth dimension, as we rearrange this dimension to generate SIs. As can be seen in Table 1, the 2D network driven by SIs with the grid size of 8×6 and 6×8, the most square-like rectangles, yielded the DSC of 0.778 and 0.777 respectively. This shows that the approach is indeed comparable to its 3D counterpart in the task. The recall scores are slightly lower than 3D U-Net in both cases, but they are compensated with higher precision scores. The grids with more disproportionate aspect ratios performed lower than more square-like grids, ranging from 0.744 to 0.770 in DSC. That was consistent in all experiments, indicating that the more square the SI, the better the performance.

In the atrial segmentation task, because the dataset contains only 100 scans, we used 4-fold cross validation, leaving more to validation set for better generalizability. In this set of experiments as well, there are two separate settings: (i) a U-Net comparison, and (ii) experimental image preprocessing for SIs.

For the first setting of the comparison of the networks, the images with the size of $512 \times 512 \times 88$ were used. The DSC of 0.893, precision of 0.898, and recall of 0.894 were achieved with the 3D U-Net. The SI generation with this size was using the grid layout of 11×8. This was not the most favorable combination for generating SIs since its aspect ratio is high, therefore, it could reach only 0.812 DSC, 0.902 precision, and 0.785 recall values.

In the second setting, the images were preprocessed. The initial preprocessing was dropping the same number of slices from either side of the depth dimension to (a) decrease the volume of the data, and (b) get one-to-one aspect ratio for the SI. That is, the scans with 88 slices in depths were clipped to have 64 slices. Simple preprocessing such as this boosted the DSC score of 2D model to 0.851. In the next step, 2D U-Net was initialized with ImageNet1k pretrained weights, pushing the DSC to 0.895 which is a 0.044 DSC jump from the base model. In the last experiment, this pretrained model was used with several sets of augmentations to see how far the model can go using a simple 2D U-Net. The augmentations were random flip, random affine, random elastic deformation, random anisotropy, and random gamma, and they are specific only for this experiment. This much aggressive augmentation pushed the DSC 0.901. This is a substantial increase from the baseline that could reach 0.812 DSC, proving that small preprocessing techniques are highly useful for the SI-based network to learn.



(a) **3D U-Net** (b) **2D U-Net**

Fig. 2: The figure shows qualitative results on 3D U-Net (on volume) and 2D U-Net (on SI) segmentation results on HECKTOR dataset sample respectively. White is the ground truth and red represents the prediction mask. Note that 3D U-Net results were cast to an SI form after its prediction for full-view comparison.

2.5 Qualitative Analysis

To understand how 3D U-Net on volumetric data and 2D U-Net on SIs perform, we conducted qualitative analysis as well. Figure 2 depicts the segmentation results for 3D (left) and 2D (right) U-Nets on HECKTOR dataset sample respectively. Note that here we are taking a sample for which both models perform similarly well. White represents the ground truth and red is the prediction mask in both images; and 3D U-Net model output was cast to the SI form after predicting on the volumetric data only to have a full-view and fair visualization. As can be seen, 2D U-Net segmenting on SIs slightly oversegments compared to 3D counterpart on volumetric data. This sometimes is preferred over undersegmenting especially since tumor is critical and it alerts both the patient and the doctor in advance.

3 Discussion

The newly proposed concept on casting the 3D problem to 2D was validated using two datasets of different modalities and tissue types. We can see that even 2D U-Net based on SIs can achieve up to par results compared to 3D U-Net. In head and neck tumor task, the model on the initial size of $144 \times 144 \times 144$ is slightly lower than the 3D network. This is because the task is challenging the variability in the CT and PET image appearance is large. Furthermore, when visually analyzed, we found that CT scans in this dataset for which the model is underperforming contain artefacts, generally in the teeth area. The artefacts can be seen in Figure 1 as well. However, when we investigated the segmentation results on multiple scans, we found that the 2D model on SIs

could better segment the tiny tumor regions located at the edges of the tumor as can be seen in Figure 2. The 3D network, although counter intuitive, missed most of the tiny tumors since it takes the volumetric data as a cube, being unable to accurately pinpoint the contouring edges.

The atrial segmentation task, being completely dissimilar from the other task, did not pose as much difficulty as the first dataset. Basic preprocessing techniques easily pushing the SI-based 2D network to have almost 9 percent increase in DSC is a good indicator of how much it further can improve. Such techniques as clipping the images from either end of the depth to generate more square-like rectangles helps the SI-based model perform better.

Thoroughly analyzing the problem, we put forth four main arguments here as to why it is preferred to use 2D over 3D networks with SIs. First, it is commonly accepted that 2D pretrained weights on large scale datasets are widely available and easily implementable. Its effect using ImageNet1k can be seen in our results. Second, there are more options for easily employable data augmentations for 2D than 3D models. Picking the right set of augmentations is crucial to model's improvement in many tasks as is widely known. Third, because of natural images being in 2D and general computer vision community, the availability of a multitude of 2D networks offers a unique look into dealing with volumetric data using SIs. Finally, self-supervised learning on 2D datasets and 2D networks is fairly much simpler, quicker, and more available, thus studying this aspect to SIs is believed to be crucial as the next step.

4 Conclusion

DL is gaining more and more audience in the medical imaging research. In its image segmentation task, we oftentimes deal with volumetric data. 3D networks, such as 3D U-Net and its variations, have been extensively studied, explored, and improved before specifically for this task. Their performance in most of the tasks are indeed impressive. In this work, we analyzed a new perspective to working with volumetric data. In the HECKTOR dataset, a simple 2D U-Net on SIs can achieve results comparable to so-claimed more powerful 3D U-Net results. Similarly, in the atrial segmentation dataset, the approach shows a promising potential, especially when it is powered with additional preprocessing techniques. We believe that there is a potential for this new perspective to dealing with 3D medical data to reach the state-of-the-art. It can be achieved with further preprocessing, using more powerful 2D networks, such as transformers or deeper CNNs, and self-supervised learning. We will look into these aspects to using super images in the future.

References

1. Ahn, B.B.: The compact 3d convolutional neural network for medical images. Standford University (2017)
2. Andrearczyk, V., Oreiller, V., Depeursinge, A.: Head and neck tumor segmentation in pet/ct. Medical Image Analysis (2021), https://www.aicrowd.com/challenges/miccai-2021-hecktor
3. Azhari, E.E.M., Hatta, M.M.M., Htike, Z.Z., Win, S.L.: Tumor detection in medical imaging: a survey. International Journal of Advanced Information Technology **4**(1), 21 (2014)
4. Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E.: An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 111–119. Springer (2017)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021), https://arxiv.org/abs/2102.04306
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016)
7. Colak, E., Kitamura, F.C., Hobbs, S.B., Wu, C.C., Lungren, M.P., Prevedello, L.M., Kalpathy-Cramer, J., Ball, R.L., Shih, G., Stein, A., et al.: The rsna pulmonary embolism ct dataset. Radiology: Artificial Intelligence **3**(2), e200254 (2021)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Domingues, I., Pereira, G., Martins, P., Duarte, H., Santos, J., Abreu, P.H.: Using deep learning techniques in medical imaging: a systematic review of applications on ct and pet. Artificial Intelligence Review **53**(6), 4093–4160 (2020)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. 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)
12. Fan, Q., Chen, C.F., Panda, R.: Can an image classifier suffice for action recognition? In: International Conference on Learning Representations (2021)
13. Feng, X., Tustison, N.J., Patel, S.H., Meyer, C.H.: Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. Frontiers in Computational Neuroscience **14**, 25 (2020). https://doi.org/10.3389/fncom.2020.00025, https://www.frontiersin.org/article/10.3389/fncom.2020.00025
14. Iantsen, A., Visvikis, D., Hatt, M.: Squeeze-and-excitation normalization for automated delineation of head and neck primary tumors in combined pet and ct images. Lecture Notes in Computer Science p. 37–43 (2021). https://doi.org/10.1007/978-3-030-67194-5_4, http://dx.doi.org/10.1007/978-3-030-67194-5_4
15. Islam, M.S., Kaabouch, N., Hu, W.C.: A survey of medical imaging techniques used for breast cancer detection. In: IEEE International Conference on Electro-Information Technology, EIT 2013. pp. 1–5. IEEE (2013)

16. Latif, J., Xiao, C., Imran, A., Tu, S.: Medical imaging using machine learning and deep learning algorithms: a review. In: 2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET). pp. 1–5. IEEE (2019)
17. Liu, S., Zhang, D., Song, Y., Peng, H., Cai, W.: Triple-crossing 2.5 d convolutional neural network for detecting neuronal arbours in 3d microscopic images. In: International Workshop on Machine Learning in Medical Imaging. pp. 185–193. Springer (2017)
18. Masood, S., Sharif, M., Masood, A., Yasmin, M., Raza, M.: A survey on medical image segmentation. Current Medical Imaging **11**(1), 3–14 (2015)
19. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
20. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
21. Patravali, J., Jain, S., Chilamkurthy, S.: 2d-3d fully convolutional neural networks for cardiac mr segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 130–139. Springer (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
23. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R.M.: A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In: International conference on medical image computing and computer-assisted intervention. pp. 520–527. Springer (2014)
24. Saeed, N., Majzoub, R.A., Sobirov, I., Yaqub, M.: An ensemble approach for patient prognosis of head and neck tumor using multimodal data (2022)
25. Saint-Esteven, A.L.G., Bogowicz, M., Konukoglu, E., Riesterer, O., Balermpas, P., Guckenberger, M., Tanadini-Lang, S., van Timmeren, J.E.: A 2.5 d convolutional neural network for hpv prediction in advanced oropharyngeal cancer. Computers in biology and medicine p. 105215 (2022)
26. Sobirov, I., Nazarov, O., Alasmawi, H., Yaqub, M.: Automatic segmentation of head and neck tumor: How powerful transformers are? arXiv preprint arXiv:2201.06251 (2022)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical Image Analysis **67**, 101832 (2021)
29. Yang, J., Huang, X., He, Y., Xu, J., Yang, C., Xu, G., Ni, B.: Reinventing 2d convolutions for 3d images. IEEE Journal of Biomedical and Health Informatics **25**(8), 3009–3018 (2021)
30. Zhang, Y., Liu, H., Hu, Q.: Transfuse: Fusing transformers and cnns for medical image segmentation. CoRR **abs/2102.08005** (2021)
31. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. CoRR **abs/1711.10684** (2017), http://arxiv.org/abs/1711.10684