Cut to the Mix: Simple Data Augmentation Outperforms Elaborate Ones in Limited Organ Segmentation Datasets

Chang Liu[1], Fuxin Fan[1], Annette Schwarz[1,2], and Andreas Maier[1]

[1] Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany
[2] Siemens Healthineers, 91027 Forchheim, Germany

Abstract. Multi-organ segmentation is a widely applied clinical routine and automated organ segmentation tools dramatically improve the pipeline of the radiologists. Recently, deep learning (DL) based segmentation models have shown the capacity to accomplish such a task. However, the training of the segmentation networks requires large amount of data with manual annotations, which is a major concern due to the data scarcity from clinic. Working with limited data is still common for researches on novel imaging modalities. To enhance the effectiveness of DL models trained with limited data, data augmentation (DA) is a crucial regularization technique. Traditional DA (TDA) strategies focus on basic intra-image operations, i.e. generating images with different orientations and intensity distributions. In contrast, the inter-image and object-level DA operations are able to create new images from separate individuals. However, such DA strategies are not well explored on the task of multi-organ segmentation. In this paper, we investigated four possible inter-image DA strategies: CutMix, CarveMix, ObjectAug and AnatoMix, on two organ segmentation datasets. The result shows that CutMix, CarveMix and AnatoMix can improve the average dice score by 4.9, 2.0 and 1.9, compared with the state-of-the-art nnUNet without DA strategies. These results can be further improved by adding TDA strategies. It is revealed in our experiments that CutMix is a robust but simple DA strategy to drive up the segmentation performance for multi-organ segmentation, even when CutMix produces intuitively 'wrong' images. Our implementation is publicly available at https://github.com/Rebooorn/mosDAtoolkit for future benchmarks.

Keywords: multi-organ segmentation · data augmentation.

1 Introduction

In clinics, multi-organ segmentation is a common routine for radiologists in order to perform a multitude of treatments or therapies, i.e. for the treatment planning for radiation therapy [9]. However, manual delineation of human organs in medical images is a time- and effort-consuming task and automated

multi-organ segmentation is thus expected. In recent years, the emerging deep learning (DL) based models have shown strong performance on the task of organ segmentation in some imaging modalities, such as computed tomography (CT) and magnetic resonance scanning (MR) [4,12]. For supervised training routines, the development of a robust DL segmentation model relies on a large-scale image dataset with manual annotation of organs. Along with the state-of-the-art organ segmentation models, many large-scale segmentation datasets are publicly available [8,5,12]. However, for research on novel imaging modalities such as dual energy computed tomography (DECT), it remains difficult to gather enough images. For preliminary researches, such as to investigate whether certain imaging modalities will benefit the DL models for organ segmentation, working with a limited segmentation dataset is still common [1].

It is a reasonable practice to extend the generalizability of limited segmentation dataset using data augmentation (DA), which is a widely known regularization method for DL [3]. Traditional data augmentation (TDA) strategies include spatial shift or scaling, and intensity scaling. These are effective for regularizing training processes with more diverse training data. Novel approaches, like inter-image and object-level DA strategies are proposed in computer vision and medical imaging research, seeking to reach further than TDA strategies. Mixup [14] and CutMix [13] were first proposed to fuse multiple images from the training dataset for image classification tasks. The concept was then adapted into the medical imaging domain, CarveMix [16] and selfMix [17] are proposed to manipulate the tumor regions within the dataset onto background images for brain and liver tumor segmentation. PII [11] is proposed for pathological anomaly detection. ObjectAug [15] introduces object-level data augmentation using the segmentation mask of the components in the image, to apply augmentation for image classification. ClassMix [10] and ComplexMix [2] are proposed in researches of autonomous driving to merge the image and the segmentation mask to generate novel street views. AnatoMix [6] is recently proposed in particular for augmentation for multi-organ segmentation task.

To the best of our knowledge, few investigations have been done for inter-image and object-level DA strategies on multi-organ segmentation task. In this work, we present our investigation on robust DA strategies for multi-organ segmentation in limited dataset. Four established DA strategies have been re-implemented to fit the multi-organ segmentation task: CutMix, ObjectAug, CarveMix and AnatoMix.

2 Method

The aforementioned DA strategies are first re-implemented for multi-organ segmentation tasks and evaluated on two limited organ segmentation tasks, aiming to find a robust DA for multi-organ segmentation. nnUNetv2 is applied to train the segmentation networks [4].

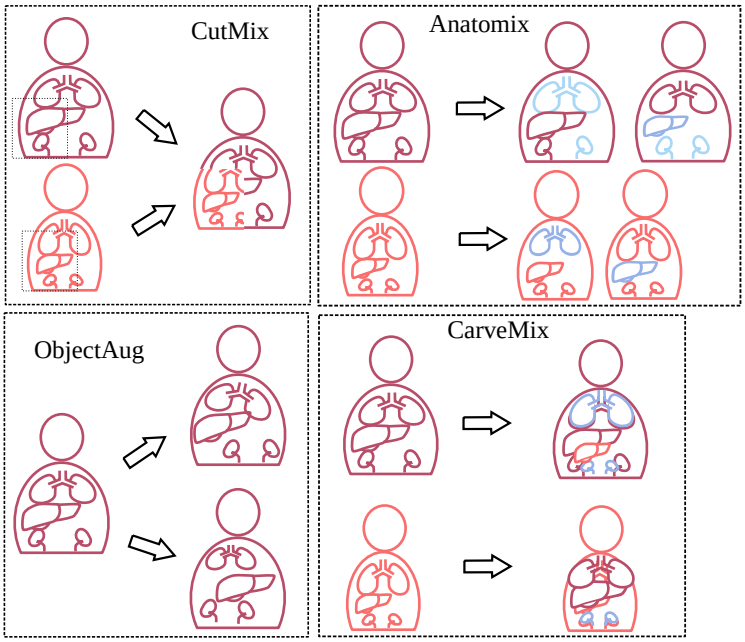

Fig. 1: Illustration of the concept of CutMix, ObjectAug, CarveMix and AnatoMix. All DA strategies are originally proposed for either image classification or tumor segmentation task. They are re-implemented for the multi-organ segmentation task and further evaluated in this work.

2.1 Inter-image and Object-level Data Augmentation

Assume the segmentation dataset contains N_d cases with $\{I_i, M_i | 0 < i \leq N_d\}$, where $I_i \in R^{D \times W \times H}$ is the volume image and the $M_i \in R^{N_{og} \times D \times W \times H}$ is the manual annotation of N_{og} organs or structures. M_i also consists of multiple channels for all organs, so $M_i = \{m_i^j | 0 < j \leq N_{og}\}$, where $m_i^j \in R^{D \times W \times H}$ is the binary mask of a specific organ in image I_i. A common operation for object manipulation in images is to overlap region of a selected image, or mask, I_s onto a background image, or mask, I_b using a binary mask m. For simplicity, in the following manuscript such fusion operation is formulated as

$$I_b \otimes (I_s, b) = I_s \cdot m + I_b \cdot (1 - m). \tag{1}$$

The four DA strategies are illustrated in Fig. 1 and their potential outputs are compared in Table. 1.

CutMix CutMix is originally proposed in computer vision research [13] and two images are fused to a new classification label. For the multi-organ segmentation task, a random bounding box mask m_{bb} is created with center at a random

Table 1: Comparison of the DA strategies re-implemented to multi-organ segmentation task.

Will the DA strategy outputs..	TDA	CutMix	ObjectAug	CarveMix	AnatoMix
have correct number of organs	Yes	No	Yes	No	Yes
have correct organ locations	Yes	No	Yes	No	Yes
cause broken organs	No	Yes	No	No	No
have artificial voxels	No	No	Yes	No	No

location with size following a ratio proportional to the image size following a Beta distribution $\beta(0.5, 0.5)$. The CutMix is then formulated as:

$$I' = I_b \otimes (I_s, m_{bb}), \tag{2}$$

$$M' = \{m_b^j \otimes (m_s^j, m_{bb})\}, \tag{3}$$

where I_b is the background image and I_s is a randomly selected image from the source dataset.

ObjectAug Like CutMix, ObjectAug is also created for the classification task. The concept is to disassemble the components within the image first and then augment each object. A background inpainting model $\theta_b(I, m_{hole})$ is thus needed because the disassemble-recombine process will come with a binary hole mask m_{hole} [7]. We implemented random scaling by 10%, random shift by 5 voxels in all dimension and random rotation of 15° for object-level augmentation, termed as G^j. The recombination process loops over each organ. The ObjectAug is then formulated as

$$I^0 = I_b', I^{j+1} = I^j \otimes (G^j(I_b), G^j(m_b^j)), \tag{4}$$

$$I' = \theta_b(I^{N_t}, m_{hole}), \tag{5}$$

$$M' = \{G^j(m_b^j)\}, \tag{6}$$

where I_b' is I_b by setting all organ pixels to background.

CarveMix CarveMix is proposed to combine the brain tumor with healthy brain region to extend the brain tumor segmentation dataset. For the multi-organ segmentation, CarveMix can be applied to each individual organ and the augmentation is then formulated as

$$I^0 = I_b, I^{j+1} = I^j \otimes (I_s, m_s^j), \tag{7}$$

$$I' = I^{N_t}, \tag{8}$$

$$M' = \{m_b^j \otimes (m_s^j, m_s^j)\}, \tag{9}$$

where I_b and I_s are the background image and a randomly selected image from the source dataset.

AnatoMix CarveMix for multi-organ segmentation will not maintain the human anatomy, in cases of organ location and organ size as can be seen in Fig. 1. To counter that, AnatoMix contains two steps: augmentation planning and organ transplant. First, the sizes of each single organ m_i^j in the dataset are analysed and a organ $m_{i'}^{j'}$ from image $I_{i'}$ with similar size will be matched for each organ m_i^j. Each organ in the background image can then be 'replaced' with similar organs in the dataset, shifted by an optimal offset S_i^j. The augmentation is formulated as

$$I^0 = I_b, I^{j+1} = I^j \otimes (S_i^j(I_{i'}), S_i^j(m_{i'}^{j'})), \tag{10}$$

$$I' = I^{N_t} \tag{11}$$

$$M' = \{m_b^j \otimes (S_i^j(m_{i'}^{j'}), S_i^j(m_{i'}^{j'}))\} \tag{12}$$

2.2 Data

Two organ segmentation datasets are used for the evaluation: the public abdominal multi-organ segmentation (AMOS) dataset and a private DECT dataset. AMOS dataset contains 300 abdominal CT volumes with segmentation of 16 organs and anatomical structures: spleen (spln), left kidney (lkdy), right kidney (rkdy), gall bladder (gbdr), esophagus (ephs), liver (livr), stomach (stmh), aorta (arta), postcava (pscv), pancreas (pcrs), right adrenal gland (rdrg), left adrenal gland (ldrg), duodenum (ddn), bladder (bldr) and prostate (prst). The DECT dataset is collected in the university hospital of Erlangen and manually annotated by a medical student, verified by a medical supervisor. The DECT dataset contains 42 CT images with segmentation of 9 abdominal organs: left kidney (lkdy), right kidney (rkdy), liver (livr), spleen (spln), left lung (llng), right lung (rlng), pancreas (pcrs), gall bladder (gbdr) and aorta (arta). For AMOS dataset, the training dataset is truncated to have only 20 images, for simulation of a limited dataset, and the full test dataset, i.e. 100 test images, are used. For DECT dataset, 20 images are used for training and 22 images for test.

In addition to the annotated organs, the two datasets also differ in the scanning regions. The AMOS dataset contains diverse scanning regions, but for DECT dataset the scanning region is almost consistent because the data comes from one single institute within a same time period.

2.3 Experimental Setting

For each DA strategy, we investigated the impact by the training dataset, the augmentation multiplier and the compatibility with the TDA strategies. For each dataset, we apply each DA strategy to augment 10, 25 and 50 times the size of the original dataset, namely 200, 500 and 1,000 images. The cases in the original dataset are not in the augmented training dataset. Then the nnUNet is trained on each augmented dataset. The nnUNet framework is selected because the training dataset is automatically resampled in every training epoch to a fixed

number of steps, so that the test performance is controlled as much as possible to only depend on the DA strategies being applied.

The DA strategies are first evaluated with no TDA to focus on the increase of generalizability. Also we present the performance of each DA strategy with the optimized TDA of nnUNet, as it is a common practice to combine such DA strategies with TDAs. The dice scores (dsc) of each organ are used for evaluation and aggregated in two ways: The macro averaged dsc aggregates the dice score of each single organ in each test sample, and the micro averaged dsc aggregates each metric from each organ in each sample, then lead to the globally averaged dsc. The micro averaged dsc is effective to indicate the general accuracy of prediction and the macro averaged dsc is more sensitive to segmentation of small objects, both are important factors for multi-organ segmentation. All experiments are done on 4 Nvidia A100 GPU (40G).

3 Results and Discussion

Table 2: Results on the AMOS dataset. 'Micro' and 'Macro' indicate two aggregation strategies. All models are tested on 100 images and trained on 20×multiplier augmented data. Three multipliers are investigated: ×10, ×25, ×50.

	Micro				Macro			
	×1	×10	×25	×50	×1	×10	×25	×50
NoTDA	88.1	–	–	–	75.9	–	–	–
+CutMix	–	89.7	90.3	**90.7**	–	78.7	79.9	**80.8**
+ObjectAug	–	46.7	50.4	44.4	–	7.4	8.9	8.1
+CarveMix	–	88.6	89.3	89.4	–	77.9	77.6	77.5
+AnatoMix	–	88.8	88.9	88.8	–	77.1	77.8	77.7
TDA	88.1	–	–	–	78.2	–	–	–
+CutMix	–	90.6	91.0	**91.1**	–	82.3	82.8	**83.0**
+ObjectAug	–	66.0	64.4	64.3	–	25.7	22.7	22.6
+CarveMix	–	89.2	89.7	89.6	–	80.0	80.7	80.2
+AnatoMix	–	89.9	89.9	89.7	–	81.5	81.4	80.8

The output of the aformentioned DA strategies are shown in Fig. 2. It can be observed that different DA strategies lead to different consistency with the original dataset. AnatoMix can produce the CT volumes with correct organ location and similar organ size. In contrast, CutMix and CarveMix rely on the similarity of both input images. When the scanning regions are greatly different, i.e. in the AMOS dataset, CutMix and CarveMix will disturb the human anatomy. For example the output volumes may have four kidneys and two livers or the upper body region will be in the lower body, as indicated by the red arrows and dashed lines in Fig. 2. On the same device, CutMix takes on average 0.3s for

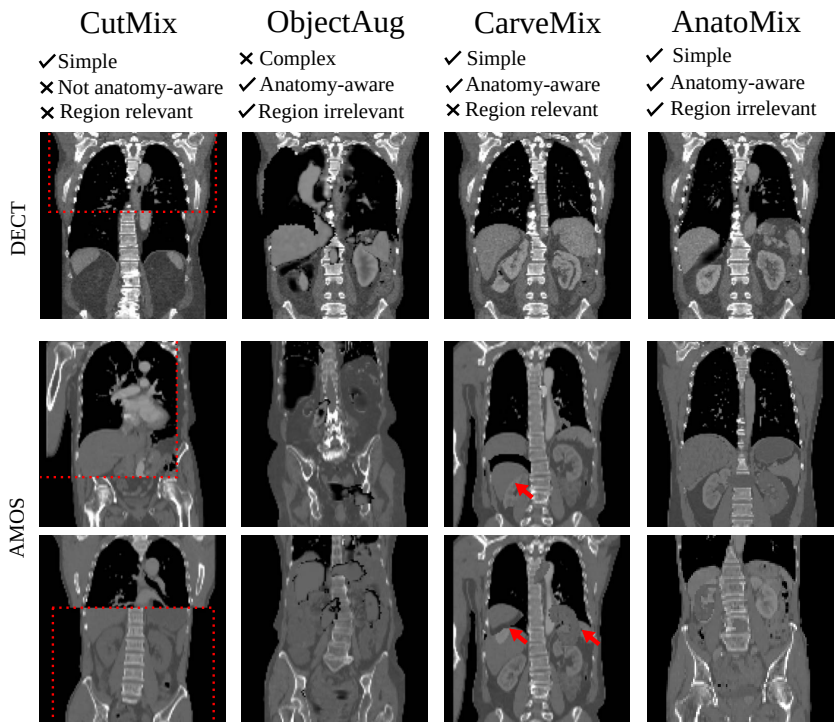

Fig. 2: Some example slices of the output volumes using the different DA strategies. The first row shows the outputs from the DECT dataset and the lower rows show the outputs from AMOS dataset. The four DA strategies lead to different compliance with the original human anatomy. Red arrows and dashed lines indicate the abnormal regions.

one output image, while it takes 15.7s for CarveMix, 20.9s for AnatoMix and 40.4s for ObjectAug. Because CutMix only combines two images by region-of-interest (ROI), it is much faster than other methods using slow operations, like background in-painting or object rotation.

The segmentation results on AMOS and DECT dataset are shown in Table. 2 and Table. 3, respectively. The detailed organ-wise results can be found in the supplementary materials. On AMOS dataset with no TDA, the applied DA strategies, except for ObjectAug, can lead to increased segmentation performance. In particular, CutMix can improve the micro averaged dsc the most by 2.6 without TDA, followed by CarveMix by 1.3 and AnatoMix by 0.8. Regarding macro averaged dsc, CutMix, CarveMix and AnatoMix each improves by 4.9, 2.0 and 1.9. Together with TDA strategies, CutMix leads to improvement of micro dsc by 3.0, followed by AnatoMix by 1.8 and CarveMix by 1.6. CutMix, CarveMix and AnatoMix each lead to the improvement of macro averaged dsc by 4.8, 2.5 and 3.2. With or without TDA, increasing the augmentation multiplier

Table 3: Results from experiments on the DECT dataset. All models are tested on 20 images and trained on 20× multiplier augmented data. Three multipliers are investigated: ×10, ×25, ×50

	Micro				Macro			
	×1	×10	×25	×50	×1	×10	×25	×50
NoTDA	96.9	–	–	–	87.7	–	–	–
+CutMix	–	**97.1**	97.0	97.0	–	**90.8**	89.1	89.3
+ObjectAug	–	76.1	77.9	76.9	–	26.4	27.5	25.7
+CarveMix	–	96.8	96.7	96.7	–	90.5	89.1	89.0
+AnatoMix	–	96.9	96.9	96.9	–	90.1	89.3	89.2
TDA	96.8	–	–	–	89.1	–	–	–
+CutMix	–	97.0	**97.1**	97.0	–	90.6	90.4	90.7
+ObjectAug	–	90.7	91.0	89.5	–	60.5	63.3	62.5
+CarveMix	–	96.9	96.8	96.9	–	90.7	90.1	90.3
+AnatoMix	–	97.0	97.0	97.0	–	90.5	90.9	**90.9**

can increase the micro and macro averaged dsc on the AMOS dataset for CutMix, but not for CarveMix and AnatoMix. The increase of macro averaged dice is higher than that for micro averaged dsc, indicating the segmentation of small organs is improved. Moreover, the increase of macro averaged dsc by CutMix is higher than that by the optimized TDA from nnUNet, and both increases are additive, leading to a joint increase of 7.0 compared with no TDA.

On the DECT dataset, the baseline performance without any DA already push towards quite high dsc, potentially because the dual channel inputs lower the difficulty of segmentation. Still, except for ObjectAug the DA strategies can slightly increase the segmentation performance. In particular without TDA, all DA strategies lead to no improvement in micro averaged dsc. In contrast, CutMix leads to improvements of 3.1 in macro averaged dsc, followed by CarveMix and AnatoMix by 2.8 and 2.4. Similarly with TDA applied, CutMix, CarveMix and AnatoMix each leads to increase by 1.6, 1.6 and 1.8. while no improvement is observed in micro averaged dsc. Different from the results on the AMOS dataset, increasing augmentation multiplier will not increase the micro or macro dsc and can even decrease the macro averaged dsc for CutMix, CarveMix and AnatoMix when no DA applied. Nevertheless, the difference between the highest and lowest macro averaged dsc is close, which indicates that DA strategies will lead to less improvement when the segmentation result is already high enough with limited datasets.

It is commonly presumed that data augmentation methods should create in-distribution data as in the original dataset. For the image data for organ segmentation, this could lead to the expectation that general characteristics of the human anatomy are preserved, like the number and relative location of organs. In other words, two livers and four kidneys in the output images should

be avoided. However, it is revealed through our research that such presumption is not always true for DL-based networks.

4 Conclusion

From our experiments, it can be concluded that the CutMix, CarveMix and AnatoMix can effectively enhance the limited segmentation datasets. In practice, it is a working strategy to combine such DA strategies with TDA to yield a joint improvement of the segmentation performance. Surprisingly, from the metric results in our experiment and the complexity of implementation, the CutMix is the best DA strategy for limited multi-organ segmentation datasets.

Acknowledgments. The authors gratefully acknowledge the HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, S., Roth, H., Dorn, S., May, M., Cavallaro, A., Lell, M., Kachelrieß, M., Oda, H., Mori, K., Maier, A.: Towards Automatic Abdominal Multi-Organ Segmentation in Dual Energy CT using Cascaded 3D Fully Convolutional Network. In: the fifth edition of The International Conference on Image Formation in X-ray Computed Tomography. pp. 395–398 (2018)
2. Chen, Y., Ouyang, X., Zhu, K., Agam, G.: Complexmix: Semi-supervised semantic segmentation via mask-based data augmentation. In: ICIP. pp. 2264–2268 (2021)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
4. 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)
5. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in NeurIPS **35**, 36722–36732 (2022)
6. Liu, C., Fan, F., Schwarz, A., Maier, A.: Anatomix: Anatomy-aware data augmentation for multi-organ segmentation. arXiv preprint arXiv:2403.03326 (2024)
7. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: The European Conference on Computer Vision (ECCV) (2018)
8. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE PAMI **44**(10), 6695–6714 (2021)
9. Nikolov, S., Blackwell, S., Zverovitch, A., et al.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv preprint arXiv:1809.04430 (2018)
10. Olsson, V., Tranheden, W., Pinto, J., Svensson, L.: Classmix: Segmentation-based data augmentation for semi-supervised learning. In: Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1369–1378 (2021)
11. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: MICCAI 2021. pp. 581–591 (2021)
12. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
13. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: ICCV (2019)
14. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
15. Zhang, J., Zhang, Y., Xu, X.: Objectaug: Object-level data augmentation for semantic image segmentation. In: IJCNN (2021)
16. Zhang, X., Liu, C., Ou, N., Zeng, X., Xiong, X., Yu, Y., Liu, Z., Ye, C.: Carvemix: A simple data augmentation method for brain lesion segmentation. In: MICCAI. pp. 196–205 (2021)
17. Zhu, Q., Wang, Y., Yin, L., Yang, J., Liao, F., Li, S.: Selfmix: a self-adaptive data augmentation method for lesion segmentation. In: MICCAI. pp. 683–692 (2022)