Soft Tissue Sarcoma Co-Segmentation in Combined MRI and PET/CT Data

Theresa Neubauer[1], Maria Wimmer[1], Astrid Berg[1], David Major[1], Dimitrios Lenis[1], Thomas Beyer[2], Jelena Saponjski[3], and Katja Bühler[1]

[1] VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Vienna, Austria
mwimmer@vrvis.at

[2] QIMP Team, Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Austria

[3] Center for Nuclear Medicine, Clinical Center of Serbia, Belgrade, Serbia

Abstract. Tumor segmentation in multimodal medical images has seen a growing trend towards deep learning based methods. Typically, studies dealing with this topic fuse multimodal image data to improve the tumor segmentation contour for a single imaging modality. However, they do not take into account that tumor characteristics are emphasized differently by each modality, which affects the tumor delineation. Thus, the tumor segmentation is modality- and task-dependent. This is especially the case for soft tissue sarcomas, where, due to necrotic tumor tissue, the segmentation differs vastly. Closing this gap, we develop a modality-specific sarcoma segmentation model that utilizes multimodal image data to improve the tumor delineation on each individual modality. We propose a simultaneous co-segmentation method, which enables multimodal feature learning through modality-specific encoder and decoder branches, and the use of resource-efficient densely connected convolutional layers. We further conduct experiments to analyze how different input modalities and encoder-decoder fusion strategies affect the segmentation result. We demonstrate the effectiveness of our approach on public soft tissue sarcoma data, which comprises MRI (T1 and T2 sequence) and PET/CT scans. The results show that our multimodal co-segmentation model provides better modality-specific tumor segmentation than models using only the PET or MRI (T1 and T2) scan as input.

Keywords: Tumor Co-segmentation · Multimodality · Deep Learning

1 Introduction

In cancer therapy, automatic tumor segmentation supports healthcare professionals as it provides a fast quantitative description of the tumor volume and location. To analyze soft tissue sarcomas in more detail, usually, complementing imaging modalities are used to depict the tumor from an anatomical or physiological perspective, such as Magnetic Resonance Imaging (MRI), Computed

Tomography (CT), or Positron Emission Tomography (PET). These modalities show different characteristics of the tumor tissue and thus provide valuable complementary information. However, depending on the imaging modality and clinical indication, the segmentation contour may look different for the same tumor.

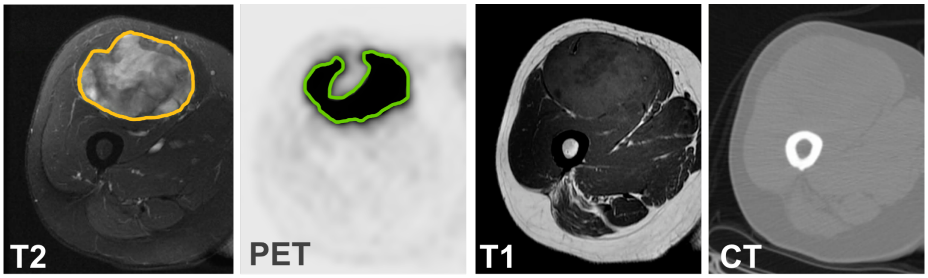

Fig. 1. Depending on the modality and the clinical intent, the segmentation for soft tissue sarcomas on the MRI T2 scan (yellow contour) and the PET scan (green contour) may look different. Figure best viewed in color.

Soft tissue sarcomas are malignant tumors that originate from various tissues, including muscular tissue, connective tissue, and nervous tissue. They predominantly occur in the extremities. Due to their large size, soft tissue sarcomas tend to form necrotic tumor areas. In MRI scans, necrosis is considered part of the tumor, but it is not visible on the PET scan as the necrosis is no longer metabolically active. Fig. 1 demonstrates the challenge of multimodal segmentation for soft tissue sarcomas on PET and MRI scans.

Deep learning based multimodal tumor segmentation methods have been proposed, e.g. for brain tumor segmentation on multi-sequence MRIs [3, 7] or lung tumor segmentation on PET/CTs [5, 12]. Current state-of-the-art networks are inspired by fully convolutional neural networks (FCNs), whereby different ways to incorporate the complementary information of multimodal image data have been presented. These multimodal segmentation studies report a better segmentation result compared to models using monomodal images. However, the main limitation of these studies is that one modality is set as the segmentation target for the final contour and thus only one modality-specific tumor volume is obtained. Contrary, in cancer therapy there are different clinical routines, which require a set of modality-specific tumor delineations from the input data.

To solve this problem for sarcomas, we aim to *simultaneously co-segment* selected modality-specific tumor volumes from the given input modalities. To the best of our knowledge, there is only the study of Zhong et al. [12], which investigates tumor co-segmentation with deep learning. They perform lung tumor segmentation on PET/CT scans, co-segmenting the modality-specific tumor in both the CT and PET scan. However, their use of two connected 3D U-Nets (one per modality), results in a very large model with more than 30M parameters.

Therefore, we introduce a resource-efficient, multimodal network for sarcoma co-segmentation, which allows the network to simultaneously segment several modality-specific tumor volumes on a subset of the input modalities. Our model benefits from (1) modality-specific encoders and decoders for multimodal feature learning, and (2) dense blocks for efficient feature re-use. We demonstrate the effectiveness of our method on public soft-tissue sarcoma data [1, 10, 11] and extensively evaluate the influence of MRI and PET/CT data for co-segmentation.

2 Method

For each patient i, $i = 1, \dots, n$, let $\mathcal{I}_i$ be a set of medical images of fixed modalities corresponding to this patient, i.e. $\mathcal{I}_i := \{I_i^m\}_{i,m}$ with I_i^m an image of patient i and modality $m \in \{T1, T2, CT, PET\}$. For every $\mathcal{I}_i$, we define the set of corresponding ground truth segmentation masks $\mathcal{M}_i := \{M_i^{m'}\}_{i,m'}$ where $m' \in \{T2, PET\}$. We then seek for a co-segmentation network that is capable of estimating the given ground truth masks $\mathcal{M}_i$, given a chosen subset of input modalities. Our proposed model is inspired by the popular U-Net [9] architecture, and the work of Jégou et al. [4], who extended the DenseNet [2] for the task of semantic segmentation. Fig. 2 gives an overview of our model, which comprises the following main parts:

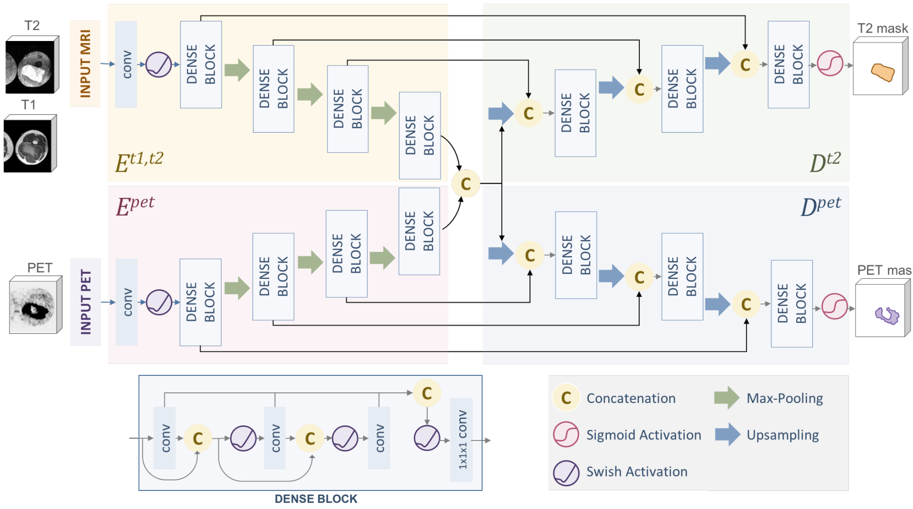

Fig. 2. We use two separated encoder branches $E^{t1,t2}$ and E^{pet} for modality-specific feature extraction and pass the concatenated latent representation to both decoders D^{t2} and D^{pet} for efficient segmentation of both tumor contours. Best viewed in color.

Modality-Specific Encoder and Decoder We use two different encoder branches $E^{t1,t2}$ and E^{pet} for MRI and PET data, respectively, to extract features for each target modality separately. In the MRI branch, we additionally use the T1 scan as a supporting modality to improve the feature extraction of the target T2 scan. The separation of the modality types in the encoder part is inspired by prior work on multimodal segmentation models [13]. Firstly, studies with multi-sequence MRIs have shown that input-level fusion leads to a significant improvement in model performance [3, 7]. Secondly, for studies dealing with complementary modalities such as PET/CT, modality-specific encoder branches are widely used [5, 12].

Each encoder, $E^{t1,t2}$ and E^{pet}, consists of an initial $3 \times 3 \times 3$ convolution layer with 48 filters, followed by four dense blocks. After each block, the resulting feature map is then downsampled using a max-pooling layer with pooling size $2 \times 2 \times 2$ and stride 2, which halves the resolution of the feature maps. To account for the low spatial resolution of the z-axis of the MRI scans, we propose to perform $2 \times 2 \times 1$ pooling after the second dense block instead.

We concatenate the latent representation of $E^{t1,t2}$ and E^{pet} and pass the feature maps to both decoders D^{t2} and D^{pet}. Each dense block in each decoder D^{t2} and D^{pet} receives the feature maps of the dense block at the same resolution level from the corresponding encoder $E^{t1,t2}$ and E^{pet}, respectively. In the following we refer to our proposed model as $E^{t1,t2}E^{pet}$-$D^{t2}D^{pet}$.

Dense Blocks Each dense block consists of three repeated $3 \times 3 \times 3$ convolution layers and Swish [8] activations. This iterative concatenation of feature maps leads to feature re-use, which in turn reduces the number of parameters [2]. The number of filters of all convolution layers in a block is increased with each block level, learning 12, 28, 44, or 60 filters, respectively. In contrast to Jégou et al. [4], we removed the batch normalization layers, since we use a batch size of one. We also removed the dropout layers, because they did not lead to performance improvements. At the end of the dense block, the feature maps of all convolution layers are then concatenated and reduced by a factor of 0.5 using a $1 \times 1 \times 1$ convolution layer to further decrease the number of model parameters.

Loss function To account for both tumor masks in our co-segmentation model during training, we calculate the dice losses individually for each mask in $\mathcal{M}_i$ and combine them as follows:

$$\mathcal{L} = - \sum_{m' \in \{T2, PET\}} \frac{2 \mid M^{m'} \cap P^{m'} \mid + \epsilon}{\mid M^{m'} \mid + \mid P^{m'} \mid + \epsilon} \tag{1}$$

whereby $M^{m'}$ and $P^{m'}$ denote the voxel set of the ground truth volume $M^{m'}$ and the predicted volume $P^{m'}$ belonging to modality $m' \in \{T2, PET\}$. The parameter ϵ is added to avoid numerical instabilities.

Variant: Shared Decoder We further introduce a lightweight variant of our model which uses only one shared decoder $D^{t2,pet}$. Here, each dense block receives the multiplied feature maps from the $E^{t1,t2}$ and E^{pet} encoder block at the same level. The fusion of feature maps by multiplication is intended to emphasize the overlapping position of the two masks. However, the feature maps of the first encoder blocks are fused by concatenation to allow for modality-specific differences in the segmentation masks. The last layer of the decoder has two output channels: one for the MRI mask M_i^{t2} and one for the PET mask M_i^{pet}. We compare both models in Section 4.

3 Experimental Setup

3.1 Dataset and Pre-processing

We evaluate our method on the soft tissue sarcoma dataset [10, 11], which is publicly available at The Cancer Imaging Archive [1]. The highly heterogeneous dataset comprises 51 patients with sarcomas in the extremities, with the data coming from different sites and scanners. For each patient, four different imaging modalities have been acquired: two paired MRI (T1 and T2) scans and a PET/CT scan. The MRI and PET/CT exams were acquired on different days, resulting in changed body positions as well as anatomical variations. The dataset already includes tumor annotations, which are delineated on the T2 scans. In addition, an experienced nuclear physician delineated the tumor contours for our study on the PET scan for radiotherapy treatment. We pre-processed the dataset as follows:

- **Co-registration**: We followed Leibfarth et al. [6] for multimodal intra-patient registration and registered the PET/CT scan with the corresponding PET contour on the T2 scan.
- **Resampling**: The in-plane pixel resolution was resampled to 0.75×0.75 mm using B-Spline interpolation, while the slice distance was kept at the original distance of the T2 scan to avoid resampling artifacts due to the low spatial resolution.
- **Crop images**: We focus on patients with tumors in their legs and cropped all scans to the leg region, resulting in 39 patients. The cropped scans have varying sizes (210×210 to 600×660) and slice numbers (15 to 49).
- **Modality-dependent intensity normalization**: We applied z-score normalization to the T1 and T2 scans. The PET scans were normalized by a transformation to standard uptake values using body-weight correction.

3.2 Network Training

We randomly divide the 39 patients into five distinct sets and perform 5-fold cross-validation. We increase the efficiency of the training using 3D patches of size $256 \times 256 \times 16$, which are randomly extracted from the image while ensuring that tumor tissue is visible on every patch. To avoid overfitting and account for the small number of training samples, we perform the following data augmentation strategies: scaling, rotation, mirroring, and elastic transformations.

We train our network using the loss function Eq. 1 and the Adam optimizer with a batch size of one. We start with an initial learning rate of $1e^{-4}$, which is reduced by a factor of 0.5 if the validation loss has not decreased for eight epochs. All convolutional kernels are initialized with he_normal. The models are implemented using Keras with Tensorflow backend and trained on an NVIDIA Titan RTX GPU (24 GB RAM).

3.3 Evaluation measures

The segmentation performance is measured calculating the overlap-based dice similarity coefficient (DSC) and distance-based average symmetric surface distance (ASSD) for each predicted mask P of modality m' and its corresponding ground truth mask $M \in \mathcal{M}_i$. Formally:

$$DSC^{m'}(M,P) = \frac{2 \mid M \cap P \mid}{\mid M \mid + \mid P \mid} \tag{2}$$

$$ASSD^{m'}(M,P) = \frac{\sum_{g_k \in M} d(g_k, M) + \sum_{p_k \in P} d(p_k, P)}{\mid M \mid + \mid P \mid} \tag{3}$$

whereby $g_k \in M$ and $p_k \in P$ denote a voxel in the ground truth volume M and predicted volume P, respectively. The Euclidean distance $d(g_k, P)$ is calculated between voxel g_k and the closest voxel in P.

4 Results and Discussion

We compare the performance of our proposed network $E^{t1,t2}E^{pet}$-$D^{t2}D^{pet}$ with different baseline models. These experiments demonstrate the influence of varying sets of input modalities as well as modality-specific encoder/decoder designs for our model. Table 1 summarizes mean DSC (in %) and ASSD (in mm) for T2 and PET segmentation separately. Visual results are shown in Fig. 3. To compare our approach to the state-of-the-art, we implement the model by Zhong et al. [12] using two parallel U-Nets: one for the T2 scan and one for the PET scan yielding the segmentation masks for the T2 and PET scan simultaneously. We followed the proposed implementation details. However, to allow for a fair comparison, we changed the patch size to our settings. Additionally, we adapt our z-axis pooling approach to the model of Zhong et al. and name it *Zhong modified.*

Single modality mask prediction: When comparing the scores for the prediction of M_i^{t2} only, we found that the lowest results are achieved when only using T2 as input. The performance increases when incorporating both T1 and T2 in the encoders, whereby the best results are obtained with a shared encoder used in model $E^{t1,t2}$-D^{t2}. These results confirm our choice for the shared MRI encoder $E^{t1,t2}$ of our proposed model. In contrast, a single PET modality is sufficient to achieve a good PET segmentation M_i^{pet}, as shown for model E^{pet}-D^{pet}.

Table 1. Performance metrics per model: Mean DSC and ASSD and their standard deviation calculated for the T2 and PET segmentation masks. All results were obtained by running a 5-fold cross-validation. Modalities used in the first encoder branch are denoted by •, and the ones in the second encoder branch are denoted by ◦.

Model	Input				Mean DSC (%)		ASSD (mm)	
	T1	T2	PET	CT	T2	PET	T2	PET
$E^{t1,t2}E^{pet}$-$D^{t2}D^{pet}$	•	•	◦		**77.2±16.5**	74.6 ± 19.0	**3.8±5.3**	4.5 ± 6.2
$E^{t1,t2}E^{pet}$-$D^{t2,pet}$	•	•	◦		75.3 ± 17.2	74.2 ± 19.9	4.5 ± 5.3	4.3 ± 5.4
$E^{t1,t2}E^{pet}$-D^{t2}	•	•	◦		76.5 ± 16.6	.	3.9 ± 4.9	.
$E^{t1,t2}E^{pet}$-D^{pet}	•	•	◦		.	74.9 ± 16.1	.	4.3 ± 5.1
E^{t2}-D^{t2}		•			65.6 ± 24.0	.	10.2 ± 10.9	.
$E^{t1,t2}$-D^{t2}	•	•			71.0 ± 23.8	.	6.5 ± 8.2	.
$E^{t1}E^{t2}$-D^{t2}	•	◦			68.3 ± 20.0	.	7.9 ± 8.9	.
E^{pet}-D^{pet}			•		.	74.3 ± 18.8	.	4.9 ± 6.4
$E^{pet,ct}$-D^{pet}			•	•	.	74.4 ± 21.6	.	5.5 ± 14.1
$E^{pet}E^{ct}$-D^{pet}			•	◦	.	**76.1±16.0**	.	3.7 ± 4.1
$E^{t2}E^{pet}$-$D^{t2}D^{pet}$		•	◦		72.1 ± 19.8	73.5 ± 20.0	4.9 ± 5.3	4.8 ± 6.2
$E^{t2}E^{pet}$-$D^{t2,pet}$		•	◦		71.6 ± 19.4	73.2 ± 19.5	5.6 ± 6.0	4.2 ± 5.1
Zhong [12]		•	◦		72.4 ± 20.0	74.1 ± 19.4	6.7 ± 8.8	4.0 ± 4.9
Zhong modified		•	◦		75.6 ± 16.2	75.2 ± 17.2	4.8 ± 6.5	**3.6±4.3**

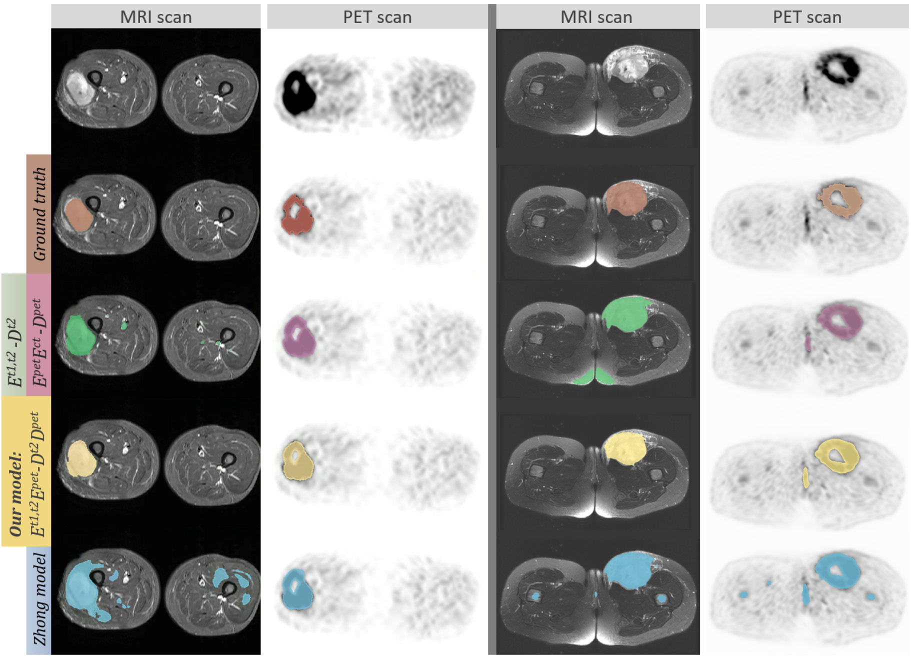

Fig. 3. Visual segmentation results of compared models on T2 and PET scan pairs. The presented samples confirm the trend from Table 1: The variation of the predicted T2 masks is higher between different models, while the impact for the PET segmentations is less apparent. Best viewed in color.

We further observed, that adding a separate encoder E^{ct} to the model resulted in the highest performance increase, yielding the best scores for predicting M_i^{pet} overall (76.1% ± 16.0% DSC, 3.7 mm ± 4.1 mm ASSD).

Encoder/Decoder design: The results in Table 1 suggest that the segmentation performance benefits from modality-specific encoders that separate anatomical and functional modalities. Comparing the models with shared $D^{t2,pet}$ and separate decoders $D^{t2}D^{pet}$, we report lower DSC scores when using the proposed shared decoder variant of our model. The performance impact is higher for T2, which is also reflected by DSC and ASSD scores.

Co-Segmentation: Looking at the sarcoma co-segmentation models, we observe that the tumor delineation on MRI T2 scans benefits from the feature co-learning with PET. This is reflected by the best overall scores (77.2% ± 16.5% DSC, 3.8 mm ± 5.3 mm ASSD) obtained with our proposed model $E^{t1,t2}E^{pet}$-$D^{t2}D^{pet}$. The same is observed with model $E^{t1,t2}E^{pet}$-D^{t2} for predicting only the T2 mask, which gives comparable results to our model. Model $E^{t1,t2}E^{pet}$-$D^{t2}D^{pet}$ outperforms the method by Zhong et al. [12] and achieves similar DSC and ASSD values to the model *Zhong modified*. However, our $D^{t2,pet}$ models (max. 3.9M) and $D^{t2}D^{pet}$ models (max. 5.5M) require - depending on the encoder - only 10-16% of the parameters of Zhong et al. (33.7M) and are therefore much more resource-efficient. When comparing the model of Zhong et al. with *Zhong modified*, it is revealing that the adaption of the pooling strategy to the anisotropic data resolution yields notable performance gains.

5 Conclusion

In this paper, we proposed a simultaneous co-segmentation model for soft tissue sarcomas, which utilizes densely connected convolutional layers for efficient multimodal feature learning. We performed an extensive evaluation, comparing various ways to incorporate multimodal data (MRI T1 and T2, CT and PET) into our model. We showed that our proposed network outperforms the state-of-the-art method for tumor co-segmentation, yielding better or comparable results for MRI T2 and PET, respectively. Moreover, our proposed co-segmentation architecture and single-modal variants reduce the number of parameters by up to 90% compared to the concurring method. These experiments show (1) improved accuracy when using multimodal data and (2) demonstrate that the choice of input modalities and encoder-decoder architecture is crucial for the segmentation result.

Acknowledgements

VRVis is funded by BMK, BMDW, Styria, SFG and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (854174) which is managed by FFG.

References

1. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013). https://doi.org/10.1007/s10278-013-9622-7
2. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: Proceedings of the IEEE CVPR. pp. 4700–4708. IEEE (2017)
3. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No New-Net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019)
4. Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In: Proceedings of the IEEE CVPR Workshops. pp. 11–19. IEEE (2017)
5. Kumar, A., Fulham, M., Feng, D., Kim, J.: Co-learning feature fusion maps from PET-CT images of lung cancer. IEEE Trans. Med. Imaging **39**(1), 204–217 (2020)
6. Leibfarth, S., Mönnich, D., Welz, S., Siegel, C., Schwenzer, N., Schmidt, H., Zips, D., Thorwarth, D.: A strategy for multimodal deformable image registration to integrate PET/MR into radiotherapy treatment planning. Acta Oncologica **52**(7), 1353–1359 (2013)
7. Myronenko, A.: 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019)
8. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015)
10. Vallières, M., Freeman, C.R., Skamene, S.R., El Naqa, I.: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. The Cancer Imaging Archive (2015). https://doi.org/10.7937/K9/TCIA.2015.7GO2GSKS
11. Vallières, M., Freeman, C.R., Skamene, S.R., El Naqa, I.: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. **60**(14), 5471–5496 (2015)
12. Zhong, Z., Kim, Y., Plichta, K., Allen, B.G., Zhou, L., Buatti, J., Wu, X.: Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. Medical Physics **46**(2), 619–633 (2019)
13. Zhou, T., Ruan, S., Canu, S.: A review: Deep learning for medical image segmentation using multi-modality fusion. Array **3-4**(10004), 1–11 (2019)