Multimodal CT and MR Segmentation of Head and Neck Organs-at-Risk

Gašper Podobnik[1](✉), Primož Strojan[2], Primož Peterlin[2],
Bulat Ibragimov[1,3], and Tomaž Vrtovec[1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
gasper.podobnik@fe.uni-lj.si
[2] Institute of Oncology Ljubljana, Ljubljana, Slovenia
[3] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Abstract. Radiotherapy (RT) is a standard treatment modality for head and neck (HaN) cancer that requires accurate segmentation of target volumes and nearby healthy organs-at-risk (OARs) to optimize radiation dose distribution. However, computed tomography (CT) imaging has low image contrast for soft tissues, making accurate segmentation of soft tissue OARs challenging. Therefore, magnetic resonance (MR) imaging has been recommended to enhance the segmentation of soft tissue OARs in the HaN region. Based on our two empirical observations that deformable registration of CT and MR images of the same patient is inherently imperfect and that concatenating such images at the input layer of a deep learning network cannot optimally exploit the information provided by the MR modality, we propose a novel modality fusion module (MFM) that learns to spatially align MR-based feature maps before fusing them with CT-based feature maps. The proposed MFM can be easily implemented into any existing multimodal backbone network. Our implementation within the nnU-Net framework shows promising results on a dataset of CT and MR image pairs from the same patients. Furthermore, the evaluation on a clinically realistic scenario with the missing MR modality shows that MFM outperforms other state-of-the-art multimodal approaches.

Keywords: Multimodal segmentation · Head and neck · Organs-at-risk · Computed tomography · Magnetic resonance · nnU-Net

1 Introduction

Head and neck (HaN) cancer is a prevalent type of cancer [3] with a yearly incidence of above 1 million cases and prevalence of above 4 million cases worldwide, accounting for around 5% of all cancer sites [17]. Radiotherapy (RT) is a standard

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_71.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 745–755, 2023.
https://doi.org/10.1007/978-3-031-43901-8_71

treatment modality for HaN cancer, which aims to deliver high doses of radiation to cancerous cells while sparing nearby healthy organs-at-risk (OARs) [21]. To optimize radiation dose distribution, accurate three-dimensional (3D) segmentation of target volumes and OARs is required. Computed tomography (CT) is the primary imaging modality used for RT planning due to its ability to provide information about electron density, however, its low image contrast for soft tissues, including tumors, makes accurate segmentation of soft tissue OARs challenging. Therefore, the integration of complementary imaging modalities, such as magnetic resonance (MR), has been strongly recommended in clinical practice to enhance the segmentation of several soft tissue OARs in the HaN region [1]. This naturally poses a question of whether automatic OAR segmentation can benefit from the MR image modality. Our study therefore aims to evaluate the impact of MR integration on the quality and robustness of automatic OAR segmentation in the HaN region, therefore contributing to the growing body of research on multimodal methods for medical image analysis.

Related Work. A literature review by Zhang et al. [24] divides deep learning (DL)-based multimodal segmentation methods into three fusion strategy groups: *early*, *late* and *hybrid* (also named *layer*) fusion. The first two groups of methods are most commonly applied; early fusion comprises simple concatenation of modalities along the channel dimension before feeding them into the deep neural network. Additionally, concatenating feature maps (FMs) from separate modality encoders can also be considered as early fusion [7]. Late fusion, on the other hand, employs separate branches for each input modality and then fuses the output features by either plain concatenation or by weighing the contributions of separate branches at the decision level. For example, Zhang et al. [23] proposed an attention mechanism to fuse FMs from two separate U-Nets that accepted contrast-enhanced arterial and venous phase CT images. The third group, hybrid fusion, aims to combine the strengths of early and late fusion [24] by employing two or more separate encoders (i.e. one for each modality) and a single decoder, where features from different resolution levels of the encoder are fused and fed into the decoder that produces the final full-resolution segmentation. Such hybrid or multi-level fusion along with the adaptive fusion method represents the current trend in computer vision [24], with the self-supervised model adaptation method as a prime example [18]. One important aspect is also the missing modality scenario, meaning that the multimodal model should produce satisfactory results even if only one input modality is available. Nevertheless, the optimal fusion strategy remains an open question in need of further exploration. Similar conclusions were reached in a review of multimodal segmentation methods in the medical imaging community by Zhou et al. [25]. Most methods implement either early or late fusion, however, the layer fusion strategy was identified as a better choice, since dense connections among layers can exploit more complex and complementary information to enhance training. The highlight is HyperDenseNet, a dual-path 3D network proposed by Dolz et al. [4] that employs dense connections between two convolutional paths, and achieves improvements compared to other fusion strategies and single modality variants. However, other studies have shown that the best fusion strategy depends on the specific nature of the problem, e.g. Yan et al. [22] demonstrated that the late fusion outperforms the other two approaches for the longitudinal detection of diabetic retinopathy. Relevant to the field of multimodal segmentation are also developments on unpaired multimodal segmentation, where cross-modality learning is employed to take advantage of different image modalities covering the same anatomy, but without the constraint to collect images from the same patients [5,10,19]. Although the methodologies comprising CycleGANs and/or multiple segmentation networks [10,19] seem promising, they can be excessively complex for the task of HaN OAR segmentation where both CT and MR image modalities from the same patient are often available. Consequently, our primary focus is the paired multimodal segmentation problem, including the missing modality scenario.

Motivation. When segmenting OARs in the HaN region for the purpose of RT planning, a multimodal segmentation model that can leverage the information from CT and MR images of the same patient might be beneficial compared to separate single-modal models. Firstly, as intuition suggests, such a model would rely on the CT image for bone structures and on the MR image for soft tissues, and therefore improve the overall segmentation quality by exploiting the complementary information from both modalities. Secondly, a multimodal model would facilitate cross-modality learning by extracting knowledge from one and applying that knowledge to the other modality, potentially improving the segmentation accuracy. Several studies indicated that such an approach is feasible, for example, for improving video classification by training a model on an auxiliary audio reconstruction task [12], or for audio-based detection by using the multimodal knowledge distillation concept, where teacher networks trained on RGB, depth and thermal images improve a student network trained only on audio data [20]. Finally, from the DL infrastructure maintenance perspective, it is easier to maintain a single model that can handle both modalities than two separate models for each modality. However, clinical practice differs considerably from theory, meaning that a number of considerations must be taken into account. Firstly, although MR image acquisition is recommended, it is not always feasible due to time constraints, scanner occupancy and financial aspects. Consequently, automatic OAR multimodal segmentation is required to handle the missing modality scenario, and provide a similar segmentation quality as a single-modality system. Secondly, because CT and MR images are not acquired simultaneously and with the same acquisition parameters (e.g. resolution), there is an inherent misalignment between both modalities. This can be mitigated with image registration, but not completely, mainly due to different patient positioning that especially affects the deformation of soft tissues, and various modality-specific artifacts (e.g. motion, implants, partial volume effect, etc.).

Contributions. To tackle these considerations, we propose a mechanism named modality fusion module (MFM) that can generally be applied to any network architecture that learns features from multiple modalities, and shows promising performance also in the missing modality scenario. The advantages of the proposed MFM are the following: 1) it enables the spatial alignment of FMs from one with FMs from the other modality to further reduce errors that persist after deformable registration of input images, and enrich the FMs to improve the final OAR segmentation, 2) it significantly improves the performance of the missing modality scenario compared to other baseline fusion approaches, and 3) it performs well also on single modality out-of-distribution data, therefore facilitating cross-modality learning and contributing to better model generalizability.

2 Methods

Backbone Architecture. Our chosen backbone network is based on nnU-Net, a publicly available framework for DL-based segmentation [8] that builds on the U-Net architecture [16], adds self-configurable pre-processing, augmentation and post-processing, and employs efficient training strategies. However, nnU-Net, which uses an early fusion strategy by concatenating input images or patches before feeding them to the first network layer, may not be the optimal strategy for multimodal segmentation. Recent studies have shown that this approach does not allow the network to learn meaningful high-level features from each modality before their fusion, resulting in only simple relationships between intensities from each input modality [4,23]. This is particularly problematic when fusing CT and MR images, which differ in several aspects, such as the type and location of artifacts, acquisition parameters, and visibility of soft tissues and bone structures. While MR images can help to improve the delineations of OARs that are poorly visible in CT images, the primary delineation is always performed on CT images with the help of registered MR images. An important repercussion is that image registration errors propagate into OAR delineations, which is particularly salient in the HaN region. To address these challenges, we propose an upgraded nnU-Net network with two separate encoders, one for each modality, and a common decoder that fuses FMs using the proposed MFM that learns to infer affine transformation parameters in a single forward pass. This approach efficiently *pseudo-registers* FMs from the MR encoder with those from the CT encoder, mitigating the effects of registration errors caused by non-rigid deformation of OARs and imaging artifacts.

Modality Fusion Module. The proposed MFM draws inspiration from the work of Jaderberg et al. [9], who introduced a spatial transformer network (STN) that learns to infer transformation parameters in a single forward pass, and then uses them to transform images and/or FMs. The fundamental idea is that STN can learn meaningful features that are spatially invariant to characteristics of the input data, without the need for extra supervision, thereby enhancing task

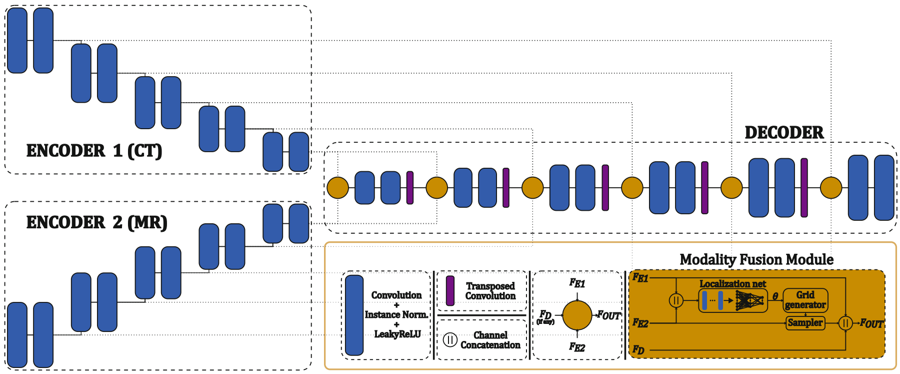


Fig. 1. The proposed backbone architecture is based on nnU-Net but with separate encoders for the computed tomography (CT) and magnetic resonance (MR) image, and with the proposed modality fusion module.

performance. While it was demonstrated that complete spatial invariance cannot be achieved with STNs [6], the work of Jaderberg et al. is crucial in showing that STNs can be implemented as differentiable modules, enabling the loss to be propagated through the sampling (interpolation) mechanism. The same underlying principle of STNs has also been leveraged in *optical flow* and its derivative work *semantic flow*, where the flow alignment module was proposed to resample low-resolution FMs and align them with high-resolution FMs [2]. We capitalize on the same principle to register FMs from MR images to those from CT images. Notably, MFM is different from semantic flow, as it takes two FMs of the same resolution but from different modalities, and aligns FMs from the auxiliary modality to FMs of the primary modality. We propose to use MFM at each resolution level of the nnU-Net backbone, which is schematically presented in Fig. 1, and consists of three blocks: *localization network*, *grid generator* and *sampler*. The localization network is a regressor network that accepts concatenated FMs from both encoders and applies four blocks of strided convolutions followed by the ReLU activation to reduce their spatial dimensions. The final FMs are flattened and fed into a simple two-layer fully connected network, which outputs 12 affine 3D transformation parameters that are then passed to the grid generator. The generated sampling grid is used by the sampler to resample FMs from the second encoder, which are then concatenated with the untouched FMs from the first encoder and the decoder (right before the bottleneck, only the first two are concatenated, as there are no decoder FMs at that level). Both the grid generator and the sampler and readily implemented in the PyTorch library [9], and because they are both differentiable, no special optimization is needed for the localization network, allowing localization parameters to be optimized with the main (segmentation) loss function. Since there is no additional supervision that would assure perfect registration, we refer to this process as *pseudo-registration*. The purpose of this architecture is to align FMs from both modalities and improve their fusion, leading to better segmentation results.

Baseline Comparison. We evaluate the performance of the proposed MFM nnU-Net against three baseline networks: 1) a single modality nnU-Net trained only on CT images, 2) a nnU-Net trained on concatenated CT and MR image pairs, and 3) a model with separate encoders for both modalities, but with a simple concatenation along the channel axis instead of the proposed MFM. In addition, we compare our model with the state-of-the-art *modality-aware mutual learning nnU-Net* (MAML) that was presented at MICCAI 2021 [23].

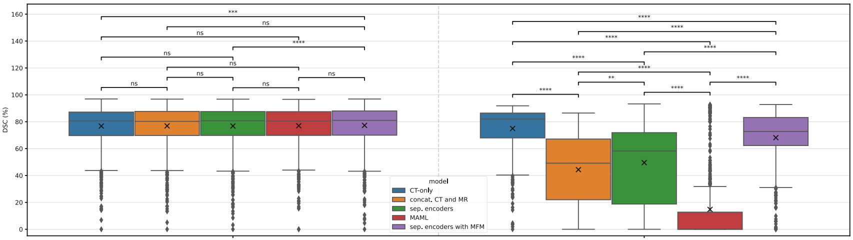
3 Experiments and Results

Image Datasets. The proposed methodology was evaluated on two publicly available datasets: our recently released HaN-Seg dataset [14] and the PDDCA dataset [15]. The HaN-Seg dataset comprises CT and T1-weighted MR images of 56 patients, which were deformably registered with the SimpleElastix registration tool, and corresponding curated manual delineations of 30 OARs (for details, please refer to [14]). Although only a subset of images is publicly available[1] due to the ongoing HaN-Seg challenge[2], both the publicly available training as well as the privately withheld test images were used in our 4-fold cross-validation experiments. On the other hand, to evaluate the generalization ability of our method, we also conducted experiments on the CT-only PDDCA dataset (for details, please refer to [15]), from which we collected 15 images from the off- and on-site test sets of the corresponding challenge for our evaluation. As this dataset is widely used for evaluating the performance of automatic HaN OAR segmentation methods, it serves as a valuable benchmark for comparison with other state-of-the-art methods. Note that none of the images from the CT-only PDDCA dataset were used for training, and as our model expects two inputs, we substituted the missing MR modality with an empty matrix (i.e. zeros).

Implementation Details. All models were trained for all OARs using the `3d_fullres` configuration of nnU-Net, with the only modification that we reduced rotation around the axial axis and disabled image flipping along the sagittal plane, which eliminated segmentation errors that were previously observed for the paired (left and right) OARs. The same modification was also used with the MAML model. To ensure a fair model comparison, we set the number of filters in the encoder of the single modality baseline model to match the number of filters of the entry-level concatenation encoder. We also halved the number of filters in networks that have separate encoders so that the overall number of parameters in the proposed model and the baselines remains approximately the same (excluding the parameters in the localization part of MFM

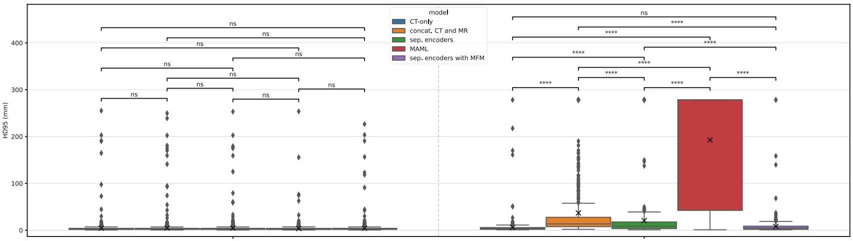
[1] https://doi.org/10.5281/zenodo.7442914.

[2] https://hanseg2023.grand-challenge.org.



dataset	HaN-Seg dataset (CT and MR images)					PDDCA dataset (only CT images)				
mean	76.1	76.2	76.2	76.3	**76.7**	**74.7**	44.1	49.3	14.7	67.8
std	11.7	**11.3**	11.4	11.4	11.4	**16.6**	22.1	24.2	27.9	19.2
min	46.1	49.1	46.8	**49.6**	46.6	**36.4**	3.5	9.5	0.0	26.5
median	77.3	76.8	76.7	76.4	**77.6**	**80.8**	44.7	52.2	4.5	74.5

Fig. 2. The results in terms of the Dice similarity coefficient (DSC) for the HaN-Seg (left) and PDDCA (right) dataset.



dataset	HaN-Seg dataset (CT and MR images)					PDDCA dataset (only CT images)				
mean	4.162	4.736	4.447	**3.696**	4.062	**6.016**	37.699	21.019	193.285	8.175
std	1.95	2.172	2.189	**1.797**	1.914	**3.164**	40.322	19.406	94.201	5.243
min	1.414	**1.368**	1.427	1.381	1.528	2.266	8.016	3.856	4.944	**2.198**
median	3.751	4.624	3.704	**3.377**	3.529	**5.107**	28.37	20.272	210.629	7.53

Fig. 3. The results in terms of the 95^{th}-percentile Hausdorff distance (HD_{95}) for the HaN-Seg (left) and PDDCA (right) dataset. (Note: Infinite values of HD_{95} were replaced with a maximal value over all data.)

block). Note that the MAML model, which is composed of two U-Nets, had a considerably higher number of parameters. To address the challenge of a relatively small dataset, we adopted a 4-fold cross-validation strategy without using any external training images. All models were trained until convergence, i.e. when the validation loss plateaued, and we selected the model with the best validation loss for inference.

Results. The quality of the obtained OAR segmentation masks was evaluated by computing the Dice similarity coefficient (DSC) and the 95^{th}-percentile Hausdorff distance (HD_{95}) against reference manual delineations, and the results for all OARs are presented in Figs. 2 and 3, respectively. Since not all images contain all 30 OARs (due to a different field-of-view), we first calculated the mean metric for each OAR and then the overall mean across all OARs to ensure that the contributions were equally weighted. We also performed analysis of statistical significance by applying paired sample t-tests with the Bonferroni correction, presented with bars on top of the box plots (non-significant: ns ($p > 0.05$), significant: $*$ ($0.01 < p < 0.05$), $**$ ($0.001 < p < 0.01$), $***$ ($0.0001 < p < 0.001$) and $****$ ($p < 0.0001$)).

4 Discussion

In this study, we evaluated the impact on the quality and robustness of automatic OAR segmentation in the HaN region caused by the incorporation of the MR modality into the segmentation framework. We devised a mechanism named MFM and combined it with nnU-Net as our backbone segmentation network. The choice of using nnU-Net as the backbone was based on the rationale that nnU-Net already incorporates numerous state-of-the-art DL innovations proposed in recent years, and therefore validation of the proposed MFM is more challenging in comparison to simply improving a vanilla U-Net architecture, and consequently also more valuable to the research community.

Segmentation Results. The obtained results demonstrate that our model performs best in terms of DSC (Fig. 2). The resulting gains are significant compared to separate encoders and CT-only models, and were achieved with 4-fold cross-validation, therefore reducing the chance of a favorable initialization. However, DSC has been identified not to be the most appropriate metric for evaluating the clinical adequacy of segmentations, especially when the results are close to the interrater variability [13], moreover, it is not appropriate for volumetrically small structures [11]. On the other hand, distance-based metrics, such as HD_{95} (Fig. 3), are preferred as they better measure the shape consistency between the reference and predicted segmentations. Although MAML achieved the best results in terms of HD_{95}, indicating that late fusion can efficiently merge the information from both modalities, it should be noted that MAML has a considerate advantage due to having two decoders and an additional attention fusion block compared to the baseline nnU-Net with separate encoders and a single decoder. On the other hand, our approach based on separate encoders with MFM is not far behind, with a mean HD_{95} of 4.06 mm, which is more than 15% better than the early concatenation fusion. The comparison to the baseline nnU-Net with separate encoders offers the most direct evaluation of the proposed MFM. An approximate 10% improvement in HD_{95} suggests that MFM allows the network to learn more informative FMs that lead to a better overall performance.

Missing Modality Scenario. The overall good performance on the HaN-Seg dataset suggests that all models are close to the maximal performance, which is bounded by the quality of reference segmentations. However, the performance on the PDDCA dataset that consists only of CT images allows us to test how the models handle the missing modality scenario and perform on an out-of-distribution dataset, as images from this dataset were not used for training. As expected, the CT-only model performed best in its regular operating scenario, with a mean DSC of 74.7% (Fig. 2) and HD_{95} of 6.02 mm (Fig. 3). However, significant differences can be observed between multimodal methods, where the proposed model outperformed MAML and other baselines by a large margin in both metrics. The MAML model with a mean DSC of less than 15% and HD_{95} of more than 190 mm was not able to handle the missing modality scenario, whereas the MFM model performed almost as good as the CT-only model, with a mean DSC of 67.8% and HD_{95} of 8.18 mm. It should be noted that we did not employ any training strategies to improve handling of missing modalities, such as swapping input images or intensity augmentations. A possible explanation is that the proposed MFM facilitates cross-modality learning, enabling nnU-Net to extract better FMs from CT images even in such extreme scenarios.

5 Conclusions

In this study, we introduced MFM, a fusion module that aligns FMs from an auxiliary modality (e.g. MR) to FMs from the primary modality (e.g. CT). The proposed MFM is versatile, as it can be applied to any multimodal segmentation network. However, it has to be noted that it is not symmetrical, and therefore requires the user to specify the primary modality, which is typically the same as the primary modality used in manual delineation (i.e. in our case CT). We evaluated the performance of MFM combined with the nnU-Net backbone for segmentation of OARs in the HaN region, an important task in RT cancer treatment planning. The obtained results indicate that the performance of MFM is similar to other state-of-the-art methods, but it outperforms other multimodal methods in scenarios with one missing modality.

Acknowledgments. This work was supported by the Slovenian Research Agency (ARRS) under grants J2-1732, P2-0232 and P3-0307, and partially by the Novo Nordisk Foundation under grant NFF20OC0062056.

References

1. Brouwer, C.L., Steenbakkers, R.J., Bourhis, J., et al.: CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKN-PCSG, NCIC CTG, NCRI, NRG oncology and TROG consensus guidelines. Radiother. Oncol. **117**, 83–90 (2015). https://doi.org/10.1016/j.radonc.2015.07.041
2. Chen, J., Zhan, Y., Xu, Y., Pan, X.: FAFNet: fully aligned fusion network for RGBD semantic segmentation based on hierarchical semantic flows. IET Image Process **17**, 32–41 (2023). https://doi.org/10.1049/ipr2.12614
3. Chow, L.Q.M.: Head and neck cancer. N. Engl. J. Med. **382**, 60–72 (2020). https://doi.org/10.1056/NEJMRA1715715
4. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ben Ayed, I.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. IEEE Trans. Med. Imaging **38**, 1116–1126 (2019). https://doi.org/10.1109/TMI.2018.2878669

5. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE Trans. Med. Imaging **39**, 2415–2425 (2020). https://doi.org/10.1109/TMI.2019.2963882
6. Finnveden, L., Jansson, Y., Lindeberg, T.: Understanding when spatial transformer networks do not support invariance, and what to do about it. In: 25th International Conference on Pattern Recognition - ICPR 2020, Milan, Italy, pp. 3427–3434. IEEE (2020). https://doi.org/10.1109/ICPR48806.2021.9412997
7. Hu, X., Yang, K., Fei, L., Wang, K.: ACNet: attention based network to exploit complementary features for RGBD semantic segmentation. In: 26th International Conference on Image Processing - ICIP 2019, Taipei, Taiwan, pp. 1440–1444. IEEE (2019). https://doi.org/10.1109/ICIP.2019.8803025
8. 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. Nat. Methods **18**, 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z
9. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems - NIPS 2015, vol. 28. Curran Associates, Montréal, QC, Canada (2015). https://doi.org/10.48550/arxiv.1506.02025
10. Jiang, J., Rimner, A., Deasy, J.O., Veeraraghavan, H.: Unpaired cross-modality educed distillation (CMEDL) for medical image segmentation. IEEE Trans. Med. Imaging **41**, 1057–1068 (2022). https://doi.org/10.1109/TMI.2021.3132291
11. Maier-Hein, L., Reinke, A., Kozubek, M., et al.: BIAS: transparent reporting of biomedical image analysis challenges. Med. Image Anal. **66**, 101796 (2020). https://doi.org/10.1016/j.media.2020.101796
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: 28th International Conference on International Conference on Machine Learning - ICML 2011, Bellevue, WA, USA, pp. 689–696. Omnipress (2011)
13. Nikolov, S., Blackwell, S., Zverovitch, A., et al.: Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J. Med. Internet Res. **23**, e26151 (2021). https://doi.org/10.2196/26151
14. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: HaN-Seg: the head and neck organ-at-risk CT and MR segmentation dataset. Med. Phys. **50**, 1917–1927 (2023). https://doi.org/10.1002/mp.16197
15. Raudaschl, P.F., Zaffino, P., Sharp, G.C., et al.: Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. Med. Phys. **44**, 2020–2036 (2017). https://doi.org/10.1002/mp.12197
16. 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). https://doi.org/10.1007/978-3-319-24574-4_28
17. Sung, H., Ferlay, J., Siegel, R.L., et al.: Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **71**, 209–249 (2021). https://doi.org/10.3322/caac.21660
18. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. Int. J. Comput. Vis. **128**, 1239–1285 (2018). https://doi.org/10.1007/s11263-019-01188-y

19. Valindria, V.V., Pawlowski, N., Rajchl, M., et al.: Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI. In: 2018 IEEE Winter Conference on Applications of Computer Vision - WACV 2018, Lake Tahoe, NV, USA, pp. 547–556. IEEE (2018). https://doi.org/10.1109/WACV.2018.00066
20. Valverde, F.R., Hurtado, J.V., Valada, A.: There is more than meets the eye: self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR 2021, Nashville, TN, USA, pp. 11607–11616. IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.01144
21. Yan, F., Knochelmann, H.M., Morgan, P.F., et al.: The evolution of care of cancers of the head and neck region: state of the science in 2020. Cancers **12**, 1543 (2020). https://doi.org/10.3390/cancers12061543
22. Yan, Y., et al.: Longitudinal detection of diabetic retinopathy early severity grade changes using deep learning. In: Fu, H., Garvin, M.K., MacGillivray, T., Xu, Y., Zheng, Y. (eds.) OMIA 2021. LNCS, vol. 12970, pp. 11–20. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87000-3_2
23. Zhang, Y., et al.: Modality-aware mutual learning for multi-modal medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12901, pp. 589–599. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87193-2_56
24. Zhang, Y., Sidibé, D., Morel, O., Mériaudeau, F.: Deep multimodal fusion for semantic image segmentation: a survey. Image Vis. Comput. **105**, 104042 (2021). https://doi.org/10.1016/j.imavis.2020.104042
25. Zhou, T., Ruan, S., Canu, S.: A review: deep learning for medical image segmentation using multi-modality fusion. Array **3–4**, 100004 (2019). https://doi.org/10.1016/j.array.2019.100004