



Tackling the class imbalance problem of deep learning-based head and neck organ segmentation

Elias Tappeiner¹ · Martin Welk¹ · Rainer Schubert¹

Received: 5 January 2022 / Accepted: 20 April 2022 / Published online: 16 May 2022
© The Author(s) 2022, corrected publication 2022

Abstract

Purpose The segmentation of organs at risk (OAR) is a required precondition for the cancer treatment with image-guided radiation therapy. The automation of the segmentation task is therefore of high clinical relevance. Deep learning (DL)-based medical image segmentation is currently the most successful approach, but suffers from the over-presence of the background class and the anatomically given organ size difference, which is most severe in the head and neck (HAN) area.

Methods To tackle the HAN area-specific class imbalance problem, we first optimize the patch size of the currently best performing general-purpose segmentation framework, the nnU-Net, based on the introduced class imbalance measurement, and second introduce the class adaptive Dice loss to further compensate for the highly imbalanced setting.

Results Both the patch size and the loss function are parameters with direct influence on the class imbalance, and their optimization leads to a 3% increase in the Dice score and 22% reduction in the 95% Hausdorff distance compared to the baseline, finally reaching 0.8 ± 0.15 and 3.17 ± 1.7 mm for the segmentation of seven HAN organs using a single and simple neural network.

Conclusion The patch size optimization and the class adaptive Dice loss are both simply integrable in current DL-based segmentation approaches and allow to increase the performance for class imbalance segmentation tasks.

Keywords Deep learning · Segmentation · Head and neck · Class imbalance · Radiation therapy

Introduction

Cancer is after cardiovascular diseases the second most common cause of death. Among the newly diagnosed cancer incidences, statistically 3% are tumors of the head and neck (HAN) region [1]. Due to the complex anatomy of the area, characterized by a large number of small soft tissue organs, image-guided radiotherapy is the primary choice of treatment for HAN cancer. The segmentation of the organs at risk (OAR) on the planning CT scans is necessary for the radio-

therapy and the main reason of treatment delivery delays throughout the clinical pathway of the therapy. The segmentation is time-consuming, requires several highly educated medical experts and is still mainly performed manually; further observer variations are well documented [2]. Due to the time-consuming and subjective manual process, a field of research has developed around the automated segmentation of the HAN organs on medical images, with deep learning (DL) being the dominant and most successful learning-based approach [3]. Segmentation with DL can be interpreted as a voxel-wise classification problem using fully convolutional neural networks. The large difference in size of classes to be segmented can be defined as the class imbalance problem. Since the first introduction of a DL-based multi-organ HAN segmentation approach [4], it is known that the HAN area is specially affected by the class imbalance problem. In addition to the large difference in ratio of background and foreground voxels, the HAN area is characterized by large size differences between the foreground classes themselves, which is anatomically given through the differently sized organs to be segmented. As a result, the class imbalance causes a large per-

✉ Elias Tappeiner
elias.tappeiner@umit-tirol.at

Martin Welk
martin.welk@umit-tirol.at

Rainer Schubert
raienr.schubert@umit-tirol.at

¹ Department for Biomedical Computer Science and Mechatronics, UMIT—Private University for Health Sciences, Medical Informatics and Technology, Eduard-Wallnöfer-Zentrum 1, 6060 Hall in Tyrol, Tyrol, Austria

formance difference in the segmentation of large and small organs [3].

In this work, we focus on the training window or patch size as the hyper-parameter with a direct influence on the class imbalance, as most segmentation networks are, due to GPU memory constraints, trained with randomly sampled patches of the original 3D image. Hence, we introduce a measurement for the class imbalance of differently sized training patches and optimize the patch size accordingly. Additionally, we adapt the classical multi-class Dice loss formulation which does not account for missing classes within patches. Our class adaptive Dice loss formulation is robust against missing classes, which is relevant for sparse class distributions within the image dataset and for the training with smaller patch sizes. We incorporate both the class imbalance optimized patches and the class adaptive Dice loss into the currently best performing general-purpose segmentation approach, the nnU-Net framework [5], and are able to increase the performance of its baseline version. The introduced multi-class confidence analysis following the work of Li et al. [6] also reveals an increased segmentation confidence for mid-sized organs due to the class label imbalance optimized patch size.

Related work

Guo et al. [7] and Gao et al. [8] were the first to specifically address and solve the class imbalance problem of the HAN area by using several different cascaded networks. The approaches are inspired by the work of clinical experts, first segmenting large and easy anchor organs and then zooming in to segment the harder small soft tissue organs. Similarly, the authors combined a strong large organ segmentation network, a small organ localization network and specific small organ segmentation networks effectively reducing the class imbalance of each network. In their follow-up work, the FocusNetv2, Gao et al. [9] further incorporated autoencoder-based shape priors [10] and adversarial training [11] into the small organ networks, achieving a Dice score (DSC) of 0.84 and a 95% Hausdorff distance (95HD) of 2.17 mm which are the currently best reported results on the MICCAI 2015 HAN segmentation challenge reference dataset [12]. An implicit reduction in the class imbalance, especially in favor of the small organs that are often visible in just a few CT slices, is recently achieved by hybrid networks using 2D convolutions and 3D convolutions in their architecture. Chen et al. [13] used 2D convolutions for the extraction of fine edges and 3D convolutions for coarse and fine semantic features in a single UNet [14]-based architecture. Tang et al. [15] extended a 2D UNet with an additional 3D convolution-based context-aware attention path and were able to achieve state of the art using a single HAN organ segmentation network.

Differently to architectural changes of the network, adapted cost functions can also reduce the class imbalance problem of DL. Roth et al. [16] presented the first DL-based multi-organ segmentation approach of the abdominal area and applied a class weighted cross-entropy (CE) loss function. The CE is an information theoretical measurement for probability distribution differences and allows to calculate the difference between the network's voxel-wise class prediction and the ground truth. As the CE is the classical loss function for image classification, Milletari et al. [17] proposed the DSC as a volume-based overlap measurement to be used as a loss function for image segmentation. The Dice loss transforms the voxel-wise measurement into a semantic label overlap measurement and has become the state-of-the-art loss function of the field. Effectively reducing the number of measurements to the number of labels, the Dice loss also reduces the sensitivity of the loss regarding the class imbalance effect. However, the Dice loss is not able to eliminate the problem due to its intrinsic bias toward large volumes [2] as well as the remaining severe over-presence of the largest class during training. Consequently, Carole et al. [18] introduced the generalized Dice score (GDSC), which adaptively weights the DSC by the current class size. However, in a previous work [4] we showed that the GDSC introduces noise in the learning curve by the adaptive weights and missing classes in case of the common patch-based training setting. Zhu et al. [19] investigated different loss functions specifically for the imbalanced HAN area and showed the combination of the Dice loss and the focal loss [20] to outperform the plain Dice loss. Isensee et al. [5] proposed to combine the CE and dice loss to measure both the voxel-wise class predictions and the semantic label overlap and were able to show advancements in many different segmentation tasks using the combined loss function in their nnU-Net.

Another approach to analyze the class imbalance in neural networks for image segmentation is presented by Li et al. [6]. The authors found that the network output of under-represented classes tends to shift toward the decision boundary during test time, whereas well-represented classes are unaffected. As a result, the authors claim that an overfitting of the small-sized classes occurs during the training. For their analysis of the class imbalance-induced overfitting, the authors suggest to plot the logit output of the training data against the test data, which we adapt and confirm for our given multi-class setting.

Method

Dataset

For our study of the class imbalance problem in the HAN area, we utilize the MICCAI 2015 HAN auto segmentation

challenge dataset [12]. The CT images of the dataset are from the 0522 multi-institutional clinical study of the Radiation Therapy Oncology Group [21], which made the data publicly available. The study contained multiple images of 111 patients with HAN cancer of the oropharynx, the hypopharynx or the larynx. The challenge dataset includes 40 patient CT scans, with manual reference segmentations of nine structures: the left and right Parotid Gland (PG), the left and right Submandibular Gland (SG), the Optic Chiasm (OC), the Brainstem (BS), the Mandible (MA) and the left and right Optic Nerves (ON). Although the original images of the 0522 study contained OAR reference segmentations for the radiotherapy planning, no standardized segmentation protocols existed at the time of the study and the segmented structures showed considerable differences in contouring. Accordingly, the nine organs for the dataset creation are iteratively re-contoured according to current scientific standard protocols until all segmentation experts agree and the observer bias is eliminated. For the scope of the challenge, 25 specific images identified by their file names are released as training images, 10 as an offsite test set and the last 5 as an additional test set for the onsite event of the challenge. In our work, we follow the challenge protocol regarding the dataset splits and combine the off- and onsite test images to one test set for our result presentation. The Submandibular Glands are not considered in our work as not all 40 CT scans contain the corresponding reference data.

Segmentation network design

Our work is based on the 3D nnU-Net framework of Isensee et al. [5]. The authors claimed and showed that a well-parameterized UNet [14] is hard to beat for any segmentation task and accordingly defined a set of well-proven fixed parameters and additional dataset-dependent rule-based parameters for a dynamically deep UNet. The fixed parameters are the learning rate, the optimizer, the data augmentation, the number of training iterations, the patch sampling strategy, the loss function, the inference using a sliding window approach and the post-processing as a largest component analysis. The most relevant dataset-dependent parameters are the spacing and the patch size further defining the UNet architecture. The spacing is evaluated as the median of the dataset in-plane spacing and the 10th percentile of the out-plane spacing resulting in a spacing of $0.98 \times 0.98 \times 2.5$ mm. The patch size is initialized to the dataset median after resampling and iteratively enlarged, simultaneously with the depth of the UNet to fill the available GPU memory using a fixed batch size of two resulting in a patch size of $192 \times 160 \times 56$. The skeleton UNet is a basic UNet with two blocks of convolution, instance normalization [22] and nonlinearity in each resolution, starting with a channel size of 32, which is getting doubled (halved)

with each downsampling (upsampling) operation. To inject gradients deeper in the network, deep supervision with auxiliary losses are used for the upsampling layer of the encoder. For further details regarding the original 3D version of the nnU-Net, we refer to the work of Isensee et al. [5].

Class imbalance measurement

As the currently most advanced general-purpose approach for medical image segmentation, we mainly follow the 3D nnU-Net framework, but adapt the loss function and also the patch size based on our class imbalance measurement as the parameters directly influencing the class imbalance while training. Figure 1 shows the average imbalance of the organ and background volume ratios of the dataset within a training epoch for different training strategies. For the ratio measurement, the dataset is rescaled following the spacing definition of the nnU-Net. Although the histograms visually show the difference of the organ volume ratios for the presented patch size strategies, we propose to use the standard deviation σ of the class ratios as a single measurement for the class imbalance. The standard deviation of the averaged in-patch organ ratios is a single and easily interpretable value. The ratios sum up to one; accordingly, the standard deviation is the average distance to an ideally uniform distribution of in-patch organ ratios. Utilizing σ as a cost function with the patch size as parameter allows us to find the training parameter with a minimal imbalance for the given dataset.

Class adaptive dice loss

The loss function proposed by the nnU-Net is the CE+Dice loss combining probabilistic voxel-wise class predictions and label overlap measurements, which is also advised by the currently largest study of loss functions for medical image segmentation by Ma et al. [23]. The CE loss is used in its basic multi-class formulation as:

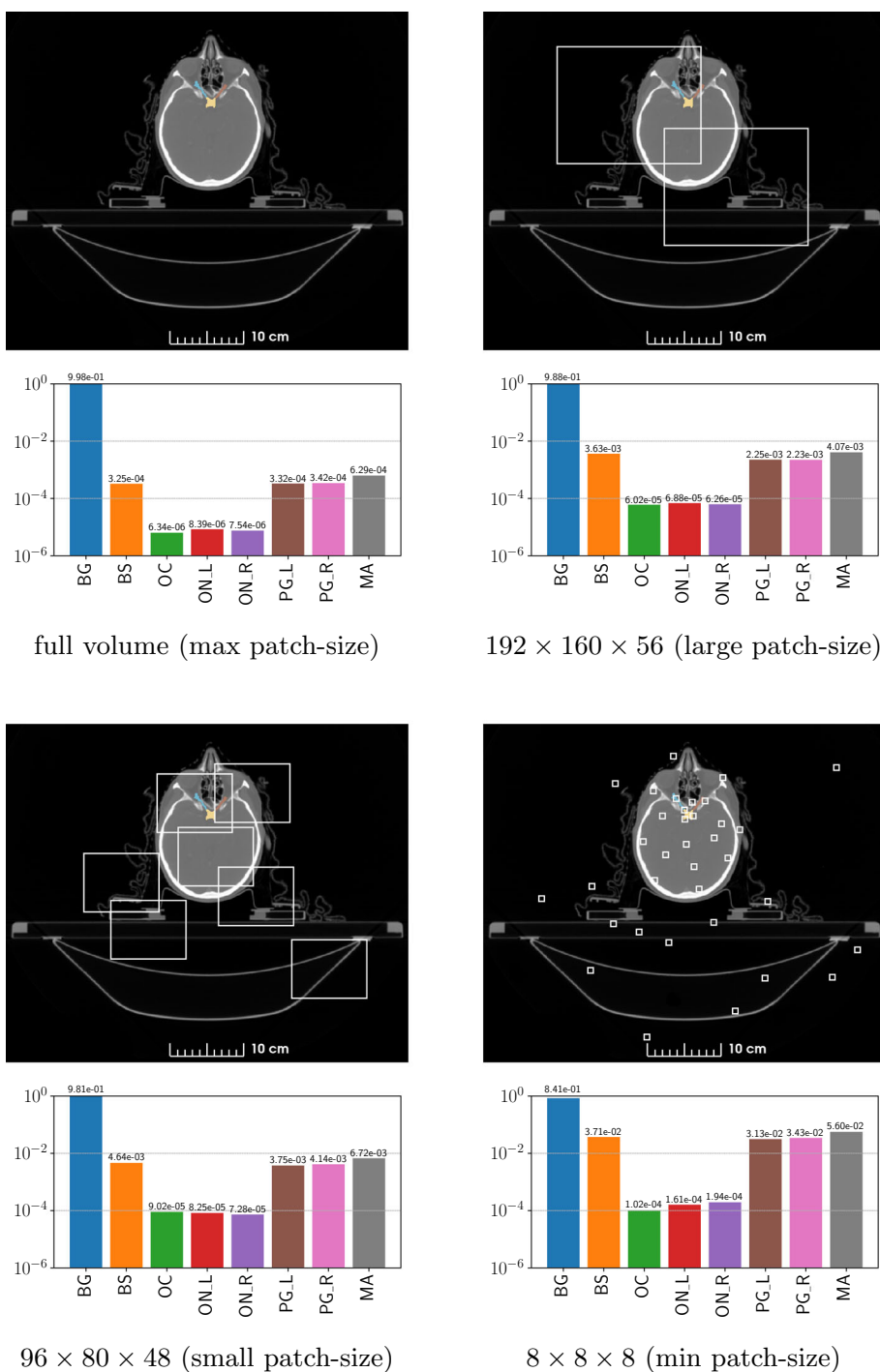
$$\text{CE}(P, G) = \frac{1}{B} - \sum_{b,c,v} G_{bcv} \log(P_{bcv}), \quad (1)$$

with P and G being the one-hot-encoded prediction and ground truth volumes, consisting of B batches, C classes and V voxels. The multi-class Dice loss using ϵ as a small value for numeric stability is defined as:

$$\text{Dice}(P, G) = \frac{1}{BC} \sum_{b,c} \frac{2 \sum_v P_{bcv} G_{bcv} + \epsilon}{\sum_v P_{bcv} + \sum_v G_{bcv} + \epsilon}. \quad (2)$$

The Dice loss formulation of the nnU-Net follows the batch Dice loss of Kodym et al. [24] with the adaptation of ignoring the background class. Contrary to the original Dice definition, Kodym et al. propose to evaluate the DSC with the batch as part of the volume instead of averaging the DSC over the

Fig. 1 Average background and organ volume imbalance ratios of seven HAN organs (from left to right: background, Brainstem, Optic Chiasm, Optic Nerve left/right, Parotid Gland left/right, Mandible) for four different patch size sampling strategies, evaluated over one training epoch



batches. Accordingly, the Dice loss formulation used in the nnU-Net is given by:

$$mnU-Dice(P, G) = \frac{1}{C - 1} \sum_{c=1}^C \frac{2 \sum_{b,v} P_{bcv} G_{bcv} + \epsilon}{\sum_{b,v} P_{bcv} + \sum_{b,v} G_{bcv} + \epsilon} \quad (3)$$

However, due to the applied patch-based training, we propose to use the class adaptive Dice loss formulation in combina-

tion with the basic CE loss. We define the class adaptive Dice loss as:

$$ca-Dice(P, G) = \frac{1}{N} \sum_{b,c} \frac{2 \sum_v P_{bcv} G_{bcv}}{\sum_v P_{bcv} + \sum_v G_{bcv} + \epsilon},$$

$$N = \sum_{b,c} \begin{cases} 0, & \text{if } \sum_v G_{bcv} = 0 \\ 1, & \text{else} \end{cases} \quad (4)$$

Differently to the original Dice loss, our definition only involves the N classes present in the sampled patch and thus evaluates to the real DSC of the sampled patch instead of considering missing classes as perfectly segmented, which biases the loss toward incorrect scores.

Results

The nnU-Net as a general-purpose segmentation framework is based on a fixed and a dataset-dependent set of parameters. The patch size defining rule of the network is based on the assumption that large windows have a more global context and hence improve the segmentation result. However, using the standard deviation σ of the organ volume ratios as a cost function to optimize the class imbalance within the patches results in smaller patch sizes than the global context maximizing patch size assumption of the nnU-Net. Our measurement of the class ratio standard deviation σ naturally shows that the class imbalance is maximal ($\sigma = 0.3301$) if a whole image approach is used and minimal if the patch size is minimal ($\sigma = 0.27146$ for a patch size of $8 \times 8 \times 8$). Figure 1 shows the organ volume ratios, including the background of the sampling process using four different patch sizes.

In our experiments, to investigate the effect of the patch size and thus the class imbalance on the segmentation quality we use the suggested patch size of the nnU-Net framework and half of the patch size in-plane and a slightly reduced size out-plane to still give the network enough context in axial direction, resulting in the small patch size $96 \times 80 \times 48$ ($\sigma = 0.32337$). We omit the full volume strategy presented in Fig. 1 as being infeasible due to its GPU memory demands as well as the minimal possible patch size only allowing a shallow U-Net with one downsampling (upsampling) layer. Additionally, we include our class adaptive Dice loss formulation (“Class adaptive dice loss” section) into the nnU-Net loss, as a robust cost function for the patch-based training of datasets with sparse class distributions as given in the HAN area. Consequently, we conduct experiments with the original nnU-Net parameters (large patch size $192 \times 160 \times 56$, nnU-Dice+CE loss) and our introduced class imbalance optimized patch size and the class adaptive loss function (ca-Dice).

Table 1 shows the average results of the networks trained on the MICCAI 2015 HAN challenge dataset [12], according to the challenge protocol. The results on the test data are evaluated using the DSC, the 95HD as well as the surface Dice (SD) as introduced by Nikolov et al. [2] combining a volume and a surface-based measurement (with surface tolerance τ identified by the authors in their observer agreement study). The bold values indicate the best results for the given measurement and values marked with stars significance (Wilcoxon signed rank test with $p < 0.05$) over the baseline. Following the work of Li et al. [6] in order to ana-

lyze a potential overfitting of the small organs we present in Fig. 2 a comparison of the output confidence distribution of the training and the test samples for the segmented organs of our experiments as violin plot. The values in each plot indicate the distance of the average confidence from the training to the test data.

Implementation details

Our implementation is based on the Monai DynUNet pipeline module,¹ a reimplementation of the dynamic UNet used in the nnU-Net framework [5] and further adapted to follow the nnU-Net parameterization. Monai is a PyTorch-based framework for deep learning in healthcare imaging.² Our models are trained on Nvidia Titan RTX GPUs with 24 GB of memory for an average of 67 hours.

Discussion

The results of our experiments in Table 1 reveal that the presented extensions to the nnU-Net framework, the patch size adjustment especially in conjunction with the class adaptive Dice loss, are favorable for the present class imbalance in the HAN area.

Reducing the patch size directly influences the class imbalance within the sampled patches. The standard deviation, introduced as a measurement for the volume ratio imbalance within a training image patch, changes from $\sigma = 0.32605$ to $\sigma = 0.32337$ using the GPU memory optimized large patch size of $192 \times 160 \times 56$ compared to the suggested class imbalance optimized small patch size of $96 \times 80 \times 48$. As visible in Fig. 1, especially the ratio of the smaller classes increases within a patch. The improvement in the class imbalance therefore reduces the bias toward the large classes during the training and effectively results in an increase in performance of 2% in terms of the DSC and a significant increase of 2% regarding SD compared to the baseline nnU-Net framework. The 95HD is significantly reduced by 0.91 mm, yielding an improvement of 22% compared the baseline.

The utilization of the class adaptive Dice loss in the loss formulation of the nnU-Net improves the segmentation results regarding the DSC and significantly the SD by another 1%. The average of the 95HD is not improved as the Optic Chiasm is not segmented in one test sample; however, all other single organ measurements show improvements over the baseline. Contrary to the standard multi-class Dice loss formulation, the class adaptive Dice loss only evaluates the classes available within each patch, whereas the standard

¹ <https://github.com/Project-MONAI/tutorials/> (accessed 21-12-21).

² <https://monai.io/> (accessed 21-12-21).

Table 1 Segmentation results on the combined on- and off-site test data of the MICCAI 2015 HAN challenge dataset [12], for the evaluated configurations in terms of DSC, 95HD and surface Dice (SD)

Config (patch size, loss)	Organ	DSC	95HD [mm]	SD
Large, nnU-Dice+CE	Brainstem	0.88 ± 0.02	3.29 ± 0.67	0.96 ± 0.03
	Optic Chiasm	0.54 ± 0.21	3.48 ± 2.02	0.84 ± 0.25
	Mandible	0.94 ± 0.01	2.13 ± 1.04	0.91 ± 0.05
	Optic Nerve_L	0.68 ± 0.11	4.91 ± 3.95	0.92 ± 0.10
	Optic Nerve_R	0.70 ± 0.08	3.10 ± 2.46	0.96 ± 0.06
	Parotid_L	0.82 ± 0.08	5.36 ± 2.45	0.91 ± 0.06
	Parotid_R	0.84 ± 0.11	6.07 ± 4.77	0.91 ± 0.10
Large, ca-Dice+CE	Brainstem	0.88 ± 0.02	3.16 ± 0.45	0.96 ± 0.06
	Optic Chiasm	0.53 ± 0.21	69.42 ± 257.44	0.85 ± 0.25
	Mandible	0.94 ± 0.01	1.86 ± 0.65	0.91 ± 0.05
	Optic Nerve_L	0.72 ± 0.08	2.82 ± 2.06	0.97 ± 0.05
	Optic Nerve_R	0.70 ± 0.07	2.21 ± 0.45	0.99 ± 0.01
	Parotid_L	0.86 ± 0.04	4.43 ± 1.62	0.94 ± 0.04
	Parotid_R	0.83 ± 0.12	5.83 ± 5.27	0.91 ± 0.12
Small, nnU-Dice+CE	Brainstem	0.88 ± 0.02	3.13 ± 0.63	0.96 ± 0.02
	Optic Chiasm	0.53 ± 0.21	3.23 ± 1.22	0.89 ± 0.13
	Mandible	0.94 ± 0.02	1.74 ± 0.75	0.92 ± 0.04
	Optic Nerve_L	0.71 ± 0.07	3.03 ± 2.08	0.96 ± 0.06
	Optic Nerve_R	0.73 ± 0.05	2.29 ± 0.48	0.98 ± 0.02
	Parotid_L	0.88 ± 0.02	4.34 ± 2.44	0.95 ± 0.03
	Parotid_R	0.88 ± 0.02	4.24 ± 1.61	0.93 ± 0.05
Small, ca-Dice+CE	Brainstem	0.88 ± 0.02	3.33 ± 0.69	0.96 ± 0.06
	Optic Chiasm	0.55 ± 0.20	3.38 ± 1.86	0.87 ± 0.19
	Mandible	0.94 ± 0.02	1.72 ± 0.69	0.92 ± 0.04
	Optic Nerve_L	0.73 ± 0.08	2.86 ± 2.13	0.97 ± 0.06
	Optic Nerve_R	0.72 ± 0.07	2.53 ± 1.45	0.98 ± 0.03
	Parotid_L	0.88 ± 0.02	4.27 ± 1.83	0.95 ± 0.02
	Parotid_R	0.87 ± 0.04	4.08 ± 1.39	0.94 ± 0.04
Large, nnU-Dice+CE	average	0.77 ± 0.17	4.05 ± 3.05	0.92 ± 0.12
Large, ca-Dice+CE	average	0.78 ± 0.16	12.82 ± 97.3*	0.93 ± 0.11*
Small, nnU-Dice+CE	average	0.79 ± 0.16	3.14 ± 1.69*	0.94 ± 0.07*
Small, ca-Dice+CE	average	0.80 ± 0.15*	3.17 ± 1.70*	0.94 ± 0.09*

Bold values indicate the best results for the respective organ in each column and values marked with stars significance (Wilcoxon signed rank test with $p < 0.05$) over the baseline

Dice loss calculates the average over all classes, distorting the average DSC depending on the current network prediction of the missing classes. The nnU-Dice which is based on the batch-Dice formulation [24], however, reduces the risk of missing classes by considering the batch dimension as part of the patch volume. The risk of missing classes within a patch depends on the volume size, the class distribution within the whole volume and, as adjustable training parameters, the patch size and the sampling strategy. As the nnU-Net framework uses a 33% random foreground oversampling strategy, the large patches and the batch-Dice formulation make the baseline nnU-Net already stable against missing classes. Nonetheless, we argue to use the class adaptive Dice, as it is robust against missing classes, especially if

the patch size is smaller and the class distribution within the volume sparse. By showing significantly improved segmentation results for all measures, our experiments support the usage of a combined small patch size and the class adaptive Dice for imbalanced segmentation problems.

Deviating from the suggestion of the original work of Li et al. [6], Fig. 2 does not show the direct network output (the logits) of the segmented classes and its corresponding decision boundaries, which is only possible for up to three classes, but the confidence distribution after the softmax normalization of the eight HAN organs to be segmented. Although no decision boundary can be depicted for more than three classes, the presentation of the organ-wise normalized confidence values allows a direct comparison of the average confidence drift

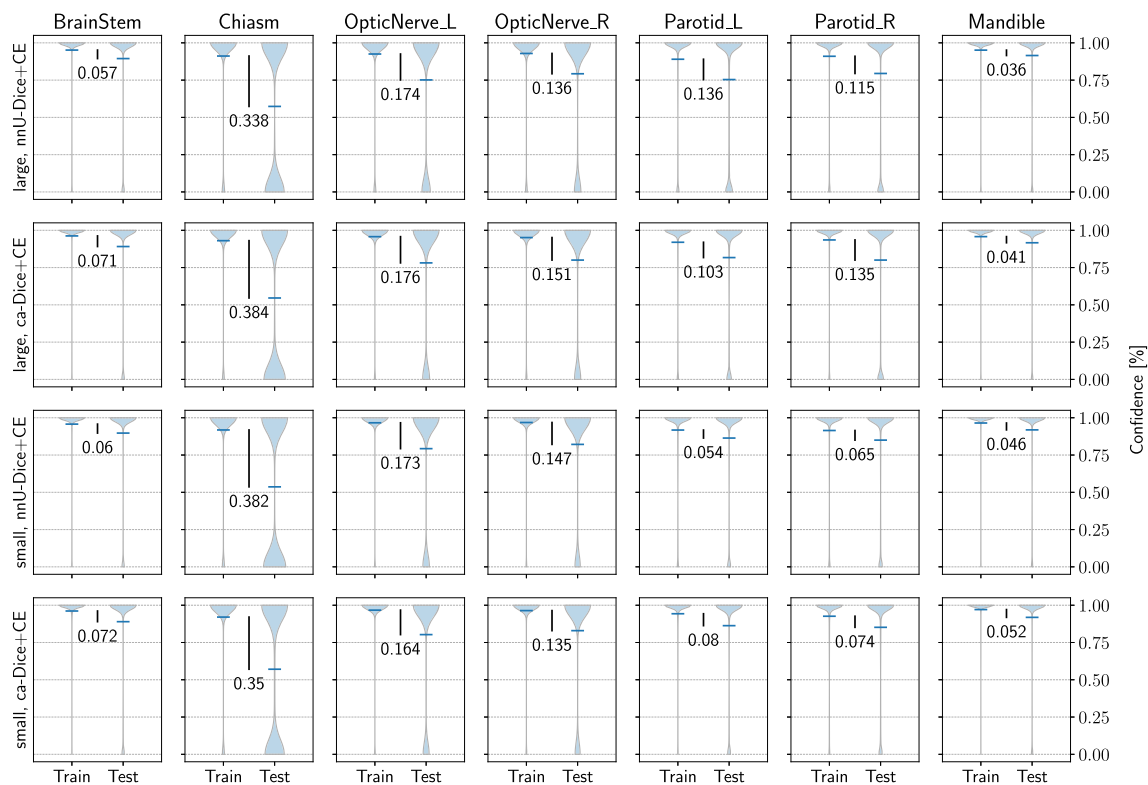


Fig. 2 Violin plot of the output confidence distribution of the training and the test samples for the segmented organs of our experiments, with the distance of the average confidence from the training to the test data indicating the potential of overfitting

from train to test time and thus the identification of overfitting. The results in Fig. 2 confirm the findings of Li et al. [6] for the class imbalance HAN area and show that the small organs (the Optic Chiasm and the Optic Nerves) are subject to larger differences in training and test time confidence and accordingly prone to overfitting. The measurements also indicate the overall performance enhancement of the ca-Dice loss over the baseline, visible in the increased average confidence values, but do not show a reduction in the overfitting of the small organs by the loss function adaption. Contrary, the experiments with the small patch size optimized to reduce the class imbalance show a clear average confidence difference reduction in the Parotid Glands. The Parotid Glands can be considered as mid-sized organs, allowing the assumption that a further reduction in the class imbalance can reduce the confidence drift for the small organs too and hence increase their final segmentation results. The assumption is also supported by the constantly small average confidence drift of the Mandible and the Brainstem being the largest organs with the largest patch ratio and consequently the least overfitting.

Finally, in Table 2 we present the segmentation results combining the small patch size and the class adaptive Dice and chronologically compare them with the segmentation results of the most important works in the field also presenting their results on the MICCAI 2015 HAN challenge

dataset. The table also indicates the number of organs and data samples used, as the original challenge protocol and its defined data splits are not followed in general.

Conclusion

In summary, in this work we present an intuitive measurement for the organ volume ratio difference, which is a central problem appearing in the DL-based segmentation of the HAN area. Based on the measurement, we optimize the patch size parameter regarding the class imbalance for a single network-based HAN segmentation architecture. Additionally, we utilize the class adaptive Dice as a robust loss function for missing classes within a training patch. Both adaptations are incorporated in the nnU-Net framework where we are able to increase the segmentation results by an additional 3% in terms of the DSC and the SD and by 22% regarding the 95HD, resulting in an average DSC of 0.8 ± 0.15 and a 95HD of 3.17 ± 1.7 mm for the segmented HAN organs, respectively.

The patch size optimization and the class adaptive Dice loss can both easily be integrated into current DL-based segmentation approaches. In future work, we want to improve the state-of-the-art performance of the recently presented

Table 2 Average DSC and 95HD on the MICCAI HAN challenge dataset

Literature	DSC	95HD [mm]	Data	Organs
Raudaschl et al.[12]	0.76	–	25/15	9 (2 partly)
Fritscher et al. [25]	0.66 ± 0.08	–	20/10	6
Tappeiner et al.[4]	0.72 ± 0.18	6.30 ± 16.2	25/15	7
Zhu et al. [19]	0.79 ± 0.05	–	38/10	9 (3 partly)
Tappeiner et al. [26]	0.75 ± 0.16	3.02 ± 1.92	25/15	7
Guo et al. [7]	0.82 ± 0.05	–	33/15	9 (3 partly)
Gao et al. [9]	0.85 ± 0.06	2.17 ± 0.93	38/10	9 (3 partly)
Nikolov et al. [2]	0.81 ± 0.05	–	(663)/15	8 (2 partly)
Chen et al. [13]	0.81 ± 0.05	–	33/10	9 (3 partly)
Tang et al. [15]	0.83 ± 0.05	–	33/15	9 (3 partly)
Our (small, ca-Dice+CE)	0.80 ± 0.15	3.17 ± 1.69	25/15	7

hybrid 2D-3D, single-network approach of Chen et al. [13] by integrating our adaptations. Single-network approaches are end-to-end trainable, less complex and therefore of higher practical interest compared to complex multi-network solutions. As an addition to the overfitting analysis, we like to investigate and combine asymmetric loss functions terms, proposed by Li et al. [6] with our ca-Dice loss to increase the distance to the decision boundaries of the small classes to further increase their test time performance.

Funding Open access funding provided by UMIT TIROL-Private Universität für Gesundheitswissenschaften und -technologie GmbH.

Availability The anonymized MICCAI 2015 HAN challenge dataset is publicly available (<http://www.imagenglab.com/newsite/pddca/>). The code of the work is available on Github (<https://github.com/ELITAP/classimbalance>).

Declarations

Conflict of interest Elias Tappeiner, Martin Welk, Rainer Schubert declare to have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed Consent Patients informed consent was given in the data originating clinical study [21].

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copy-

right holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics 2021. *CA Cancer J Clin* 71(1):7–33. <https://doi.org/10.3322/caac.21660>
2. Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, Patel Y, Meyer C, Askham H, Romera-Paredes B, Kelly C, Karthikesalingam A, Chu C, Carnell D, Boon C, D'Souza D, Moineuddin SA, Consortium DR, Montgomery H, Rees G, Suleyman M, Back T, Hughes C, Ledsam JR, Ronneberger O (2021) Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *J Med Internet Res* 23(7):26151. <https://doi.org/10.2196/26151>
3. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B (2020) Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. *Med Phys* 14(9):929–950. <https://doi.org/10.1002/mp.14320>
4. Tappeiner E, Pröll S, Hönig M, Raudaschl PF, Zaffino P, Spadea MF, Sharp GC, Schubert R, Fritscher K (2019) Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. *Int J Comput Assist Radiol Surg* 14(5):745–754. <https://doi.org/10.1007/s11548-019-01922-4>
5. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18(2):203–211. <https://doi.org/10.1038/s41592-020-01008-z>
6. Li Z, Kamnitsas K, Glocker B (2021) Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE Trans Med Imaging* 40(3):1065–1077. <https://doi.org/10.1109/TMI.2020.3046692>
7. Guo D, Jin D, Zhu Z, Ho TY, Harrison AP, Chao CH, Xiao J, Yuille A, Lin CY, Lu L (2020) Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. *Proc Conf Comput Vis Pattern Recogn* 4222–4231. <https://doi.org/10.1109/CVPR42600.2020.00428>
8. Gao Y, Huang R, Chen M, Wang Z, Deng J, Chen Y, Yang Y, Zhang J, Tao C, Li H (2019) FocusNet: imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck CT images. *Proc Conf Med Image Comput Com-*

- put Assist Interven 11766:829–838. https://doi.org/10.1007/978-3-030-32248-9_92
9. Gao Y, Huang R, Yang Y, Zhang J, Shao K, Tao C, Chen Y, Metaxas DN, Li H, Chen M (2020) FocusNetv2: imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images. *Med Image Anal* 67:1–20. <https://doi.org/10.1016/j.media.2020.101831>
 10. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, Cook SA, De Marvao A, Dawes T, O'Regan DP, Kainz B, Glocker B, Rueckert D (2018) Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. *Trans Med Imaging* 37(2):384–395. <https://doi.org/10.1109/TMI.2017.2743464>
 11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) *Adv Neural Inf Process Syst* 27:2672–2680
 12. Raudaschl PF, Zaffino P, Sharp GC, Spadea MF, Chen A, Dawant BM, Albrecht T, Gass T, Langguth C, Luthi M, Jung F, Knapp O, Wesarg S, Mannion-Haworth R, Bowes M, Ashman A, Guillard G, Brett A, Vincent G, Orbes-Arteaga M, Cardenas-Pena D, Castellanos-Dominguez G, Aghdasi N, Li Y, Berens A, Moe K, Hannaford B, Schubert R, Fritscher KD (2017) Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys* 44(5):2020–2036. <https://doi.org/10.1002/mp.12197>
 13. Chen Z, Li C, He J, Ye J, Song D, Wang S, Gu L, Qiao Y (2021) A novel hybrid convolutional neural network for accurate organ segmentation in 3d head and neck CT images. *Proc Conf Med Image Comput Assist Interven* 569–578. <https://doi.org/10.1007/978-3-030-87193-2>
 14. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. *Proc Conf Med Image Comput Assist Interven* 9351:234–241. https://doi.org/10.1007/978-3-319-24574-4_28
 15. Tang H, Liu X, Han K, Xie X, Chen X, Qian H, Liu Y, Sun S, Bai N (2021) Spatial context-aware self-attention model for multi-organ segmentation. *Proc Conf Appl Comput Vis* 938–948. <https://doi.org/10.1109/WACV48630.2021.00098>
 16. Roth HR, Oda H, Hayashi Y, Oda M, Shimizu N, Fujiwara M, Misawa K, Mori K (2017) Hierarchical 3D fully convolutional networks for multi-organ segmentation. *ArXiv preprint arXiv:1704.06382*
 17. Milletari F, Navab N, Ahmadi SA (2016) V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proc Conf 3D Vis* 565–571. <https://doi.org/10.1109/3DV.2016.79>
 18. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *Proc Int Workshop Multimodal Learn Clin Decis Support* 10553:240–248. https://doi.org/10.1007/978-3-319-67558-9_28
 19. Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, Du N, Fan W, Xie X (2019) AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys* 46(2):576–589. <https://doi.org/10.1002/mp.13300>
 20. Lin TY, Goyal P, Girshick R, He K, Dollár P (2020) Focal Loss for Dense Object Detection. *IEEE Trans Pattern Anal Mach Intell* 42(2):318–327. <https://doi.org/10.1109/TPAMI.2018.2858826>
 21. Ang KK, Zhang Q, Rosenthal DI, Nguyen-Tan PF, Sherman EJ, Weber RS, Galvin JM, Bonner JA, Harris J, El-Naggar AK, Gillison ML, Jordan RC, Konski AA, Thorstad WL, Trotti A, Beitler JJ, Garden AS, Spanos WJ, Yom SS, Axelrod RS (2014) Randomized phase III trial of concurrent accelerated radiation plus cisplatin with or without cetuximab for stage III to IV head and neck carcinoma: RTOG 0522. *J Clin Oncol* 32(27):2940–2950. <https://doi.org/10.1200/JCO.2013.53.5633>
 22. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: the missing ingredient for fast stylization. *ArXiv preprint arXiv:1607.08022*
 23. Ma J, Chen J, Ng M, Huang R, Li Y, Li C, Yang X, Martel AL (2021) Loss odyssey in medical image segmentation. *Med Image Anal* 71. <https://doi.org/10.1016/j.media.2021.102035>
 24. Kodym O, Španěl M, Herout A (2019) Segmentation of head and neck organs at risk using CNN with batch dice loss. *German Conf Pattern Recogn* 105–114. https://doi.org/10.1007/978-3-030-12939-2_8
 25. Fritscher K, Raudaschl P, Zaffino P, Spadea MF, Sharp GC, Schubert R (2016) Deep neural networks for fast segmentation of 3D medical images. *Proc Conf Med Image Comput Assist Interven* 158–165. https://doi.org/10.1007/978-3-319-46723-8_19
 26. Tappeiner E, Pröll S, Fritscher K, Welk M, Schubert R (2020) Training of head and neck segmentation networks with shape prior on small datasets. *Int J Comput Assist Radiol Surg* 15(9):1417–1425. <https://doi.org/10.1007/s11548-020-02175-2>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.