PAPER • OPEN ACCESS

Essential parameters needed for a U-Net-based segmentation of individual bones on planning CT images in the head and neck region using limited datasets for radiotherapy application

To cite this article: Ama Katseena Yawson *et al* 2024 *Phys. Med. Biol.* **69** 035008

IOP Publishing Phys. Med. Biol. 69 (2024) 035008 https://doi.org/10.1088/1361-6560/ad1996

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
26 June 2023

REVISED
30 October 2023

ACCEPTED FOR PUBLICATION
29 December 2023

PUBLISHED
24 January 2024

PAPER

Essential parameters needed for a U-Net-based segmentation of individual bones on planning CT images in the head and neck region using limited datasets for radiotherapy application

Ama Katseena Yawson[1,2,3], **Alexandra Walter**[1,4], **Nora Wolf**[1,5], **Sebastian Klüter**[2,6], **Philip Hoegen**[2,6], **Sebastian Adeberg**[7], **Jürgen Debus**[2,6,8], **Martin Frank**[4], **Oliver Jäkel**[1,2,8] and **Kristina Giske**[1,2]

[1] German Cancer Research Center (DKFZ), Division of Medical Physics in Radiation Oncology, Heidelberg, Germany
[2] Heidelberg Institute for Radiation Oncology (HIRO), National Center for Radiation Research in Oncology (NCRO), Heidelberg, Germany
[3] Heidelberg University, Medical Faculty, Heidelberg, Germany
[4] Karlsruhe Institute of Technology (KIT), Department of Mathematics, Karlsruhe, Germany
[5] Heidelberg University, Faculty of Physics and Astronomy, Heidelberg, Germany
[6] University Hospital Heidelberg, Department of Radiation Oncology, Heidelberg, Germany
[7] Marburg Ion Beam Therapy Center (MIT), Marburg, Germany
[8] Heidelberg Ion Therapy Center (HIT), Heidelberg, Germany

E-mail: a.yawson@dkfz-heidelberg.de

Keywords: head and neck cancer, radiotherapy, planning CT, bone segmentation, U-Net, nnU-Net

Abstract

Objective. The field of radiotherapy is highly marked by the lack of datasets even with the availability of public datasets. Our study uses a very limited dataset to provide insights on essential parameters needed to automatically and accurately segment individual bones on planning CT images of head and neck cancer patients. *Approach.* The study was conducted using 30 planning CT images of real patients acquired from 5 different cohorts. 15 cases from 4 cohorts were randomly selected as training and validation datasets while the remaining were used as test datasets. Four experimental sets were formulated to explore parameters such as background patch reduction, class-dependent augmentation and incorporation of a weight map on the loss function. *Main results.* Our best experimental scenario resulted in a mean Dice score of 0.93 ± 0.06 for other bones (skull, mandible, scapulae, clavicles, humeri and hyoid), 0.93 ± 0.02 for ribs and 0.88 ± 0.03 for vertebrae on 7 test cases from the same cohorts as the training datasets. We compared our proposed solution approach to a retrained nnU-Net and obtained comparable results for vertebral bones while outperforming in the correct identification of the left and right instances of ribs, scapulae, humeri and clavicles. Furthermore, we evaluated the generalization capability of our proposed model on a new cohort and the mean Dice score yielded 0.96 ± 0.10 for other bones, 0.95 ± 0.07 for ribs and 0.81 ± 0.19 for vertebrae on 8 test cases. *Significance.* With these insights, we are challenging the utilization of an automatic and accurate bone segmentation tool into the clinical routine of radiotherapy despite the limited training datasets.

1. Introduction

As of today, radiotherapy is a pillar in anti-cancer therapy and is administered to more than half of cancer patients at some point during their therapy. Radiotherapy is complex and involves many tasks that are likely to be successfully automated or at least partly supported by tools developed using machine learning techniques (Meyer *et al* 2018, Vogelius *et al* 2020). More specifically, bone delineation is an important medical imaging tool to assist clinicians with the assessment of metastatic state of cancer (Belal *et al* 2019), the facilitation of clinical decision-making for radiation treatment planning (Balagopal *et al* 2018, Kompella *et al* 2019) and many more. For example, it has been used as a supportive tool for a biomechanical image registration to enhance registration precision by Bauer *et al* (2023).

Manual contouring representing the gold standard for bone delineation is labour-intensive, time-consuming and prone to high inter-observer variability. An accurate automatic segmentation tool for delineating individual bony structures is non-trivial despite the high bone contrast offered by CT images (Minnema *et al* 2018). This is specifically linked to wide variations in the human bones in terms of shape, size and composition ranging from long to irregular bones as found in the vertebral column or the skull (Leydon *et al* 2020). Other limiting factor includes the inherently low signal-to-noise ratio, poor spatial resolution and several artifacts such as metal artifacts in CT images (Karimi *et al* 2012). Thus, individual bone segmentation is particularly difficult when relying on conventional image processing methods such as thresholding, edge detection algorithms, region growing, random walker and so on.

The emergence of deep learning based approaches largely outperforms human perception power in extracting useful information from large amounts of data such as images than conventional machine learning methods in many applications (Suzuki 2017, Xu *et al* 2018, Sahiner *et al* 2019). In this study scope, Belal *et al* (2019) published the first promising step towards a highly needed automated PET/CT-based imaging methodology in prostate cancer for 49 selected bones. Their proposed solution pipeline was a hybrid method of deep learning and shape models. Klein *et al* (2019) performed a similar study using a stand-alone deep learning method for a full-body bone tissue segmentation, without the identification of individual bones. As a follow-up study for the deep learning based approach, Schnider *et al* (2020) carried out a study for over 100 individual bones for the upper body using nnU-Net designed by Isensee *et al* (2021). Based on their findings, they discovered that their model found it very challenging to predict some bone classes which tend to be confused with each other such as the ribs and vertebrae. Therefore, further validation and evaluation are needed for stand-alone deep learning approaches before their introduction into clinical routine. On the other hand, the field of radiotherapy is highly marked with a lack of datasets even with the availability of public datasets such as the StructSeg2019 segmentation for radiotherapy planning challenge 2019 (Wahid *et al* 2023). Hence, this need raises concern on the essential parameters needed by deep learning models to yield an accurate segmentation of individual bones on planning CT images with limited training datasets.

The underlying objective of this study is to explore the capability of a U-Net-based segmentation of individual bones on planning CT images of head and neck cancer patients who received radiotherapy. Specifically, we seek to investigate different experimental cases and essential parameters which will yield an accurate prediction of all individual bones in the head and neck region with a very limited training dataset acquired from different cohorts. Furthermore, we compared the prediction capability of our best experimental scenario to benchmark segmentation tools. Finally, our model was evaluated using cohorts from the HaN-Seg challenge 2023 (Podobnik *et al* 2023) to quantify how our model generalizes to patients acquired using different protocols and acquisition parameters.

2. Materials and methods

2.1. Patient datasets

This study was conducted using 30 planning CT images of head and neck cancer patients who received radiotherapy. Image datasets were obtained from 5 different cohorts using different setup positioning devices and acquisition protocols (Stoiber *et al* 2009, Giske *et al* 2011, Stoll *et al* 2016). The data cohort included patients treated within the: German Cancer Research Center (DKFZ) (Giske *et al* 2011, Stoll *et al* 2016), Heidelberg Ion Therapy facility (HIT) (Bosch *et al* 2015), Cancer Imaging Archive (TCIA) (Ang *et al* 2014, Bosch *et al* 2015), University Clinics Heidelberg (UKHD) and HaN-Seg challenge 2023. The axial image size was 512×512 pixels, with axial slice numbers ranging from 111 to 400. The voxel spacing were in the range of $0.98 \times 0.98 \times 2$ to $1.40 \times 1.40 \times 3.3$ mm^3.

2.2. Manual annotations

Since the selected deep learning approach is a supervised learning approach, manual annotation of individual bones on each image is needed prior to the model build-up. Manual annotations were performed on all planning CT images (with the exception of the HaN-Seg challenge cohorts) by 5 observers following the anatomical guidelines as outlined by Möller (2005). Despite the existing anatomical guideline, manual annotations between the different observers was impacted by a complex set of aspects, which could be either patient-specific or protocol-specific. Therefore, the inter-observer variability was indispensable especially in the skull and vertebral bones for this task. Hence, a single observer with a high level of experience in bone segmentation task was employed to manually refine the segmentation mask of one observer whose segmentation quality was conformal to the provided anatomical guideline. In figure 1, we present the evolvement of manual annotations after an expert correction. The refinement procedure additionally involved the exclusion of teeth in both the skull and mandible; while delineation of each rib included its corresponding costal cartilage. An illustration of manually

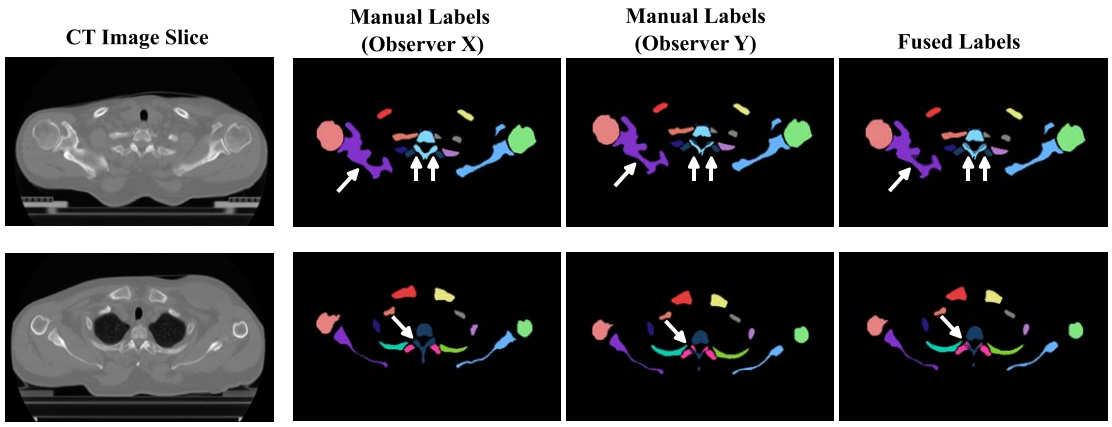

Figure 1. Evolvement of manual annotations after an expert correction for 2 instances. **Observer X** represents the observer whose manual segmentation were conformal to the provided anatomical guideline; **Observer Y** represents the corrections made by an expert; **Fused Labels** is an overlay of manual annotations of **Observer X** and **Observer Y**. The white arrows point to regions where deviations are visible.

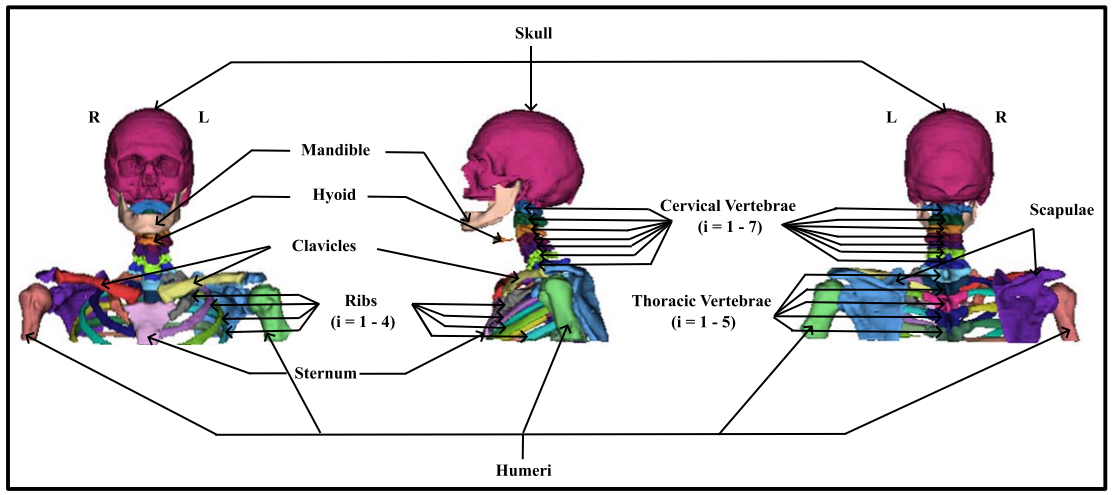

Figure 2. An illustration of manually segmented individual bones for an exemplary head and neck cancer patient. Different colours correspond to the different individual bones with their corresponding names.

segmented individual bones with their original names is depicted in figure 2. The structure set consists of 32 (excluding background) individual bones per patient after manual annotations.

2.3. Image preprocessing

Prior to the U-Net model build-up for automatic individual bone segmentation, body mask was manually outlined on each patient to remove treatment couch and regions outside the body contour were set to the Hounsfield Unit (HU) value of air (−1024 HU). Subsequently, the level of noise present in each planning CT image was corrected using an isotropic diffusion filtering; which has the ability to preserve edge information (Perona *et al* 1994). The number of individual bones was reduced to 24 using a grouping technique which assigns the left and right instances of the same bone as an individual bone for bone structures such as the scapulae, humeri, clavicles and ribs. That is, the same numeric value was assigned to left and right side of the same bone class. A sample of the default and grouped bone classes is displayed in figures 3(a) and (b) respectively.

2.4. Network description

The 3D U-Net architecture for dense volumetric segmentation was employed in our investigation (Cicek *et al* 2016). The encoding network consists of a double convolutional layer with a volumetric kernel size of $3 \times 3 \times 3$ followed by batch normalization (Ioffe and Szegedy 2015), activation function based on Leaky ReLU (Xu *et al* 2020) with a leak factor of 0.2 and a pooling layer of $2 \times 2 \times 2$. In contrast to the encoding network, the decoding

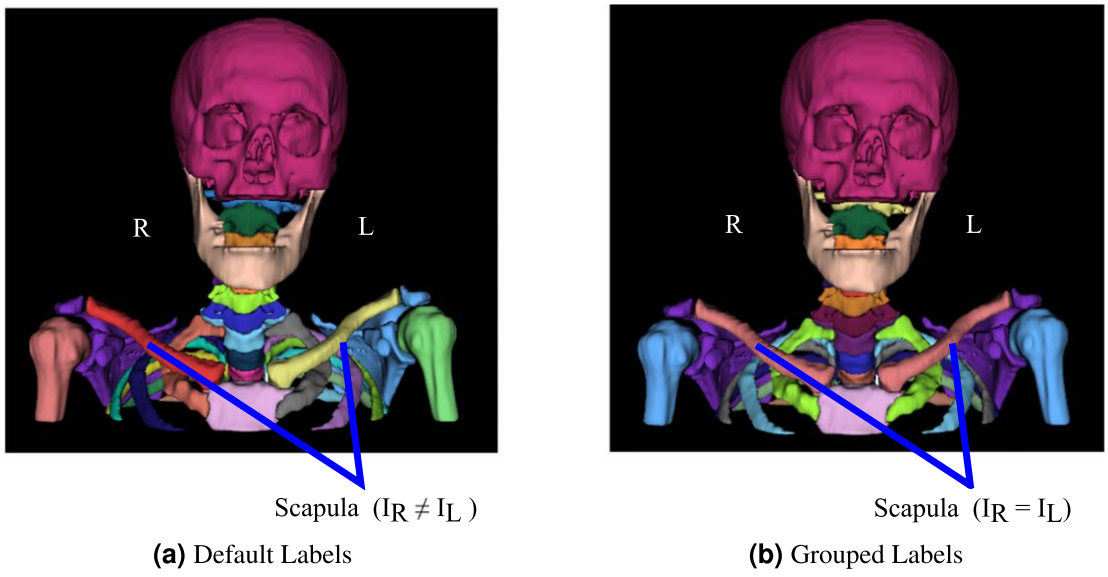

Figure 3. Grouping technique adopted to reduce individual bones. Default Labels define 32 individual bones while Grouped Labels define 24 individual bones after the grouping technique. I_R and I_L represent the numeric values of bone at the right and left instance respectively.

network is made of double up-convolutional (opposite of convolution operation) layers with a volumetric kernel size of $3 \times 3 \times 3$ each followed by Leaky ReLU. High-level feature information extracted in the encoding path was incorporated into the decoding path by concatenating them at each layer through a shortcut connection. A $1 \times 1 \times 1$ convolutional layer was performed at the output layer to reproduce the required output labels. For this selected network, adjustments were made in the base convolutional filter by initializing it to 64 and doubled whenever the network increases in depth. The number of channels assigned to the classification layer was 24, to account for the different grouped bone classes.

2.5. Model generation

Of these 30 patients, 15 cases from a mixture of DKFZ, HIT, TCIA and UKHD cohorts were randomly selected as training and validation datasets. The remaining 15 cases were divided into 2 test groups: **Test A**—7 cases from the same cohort as reflected in the training datasets and **Test B**—8 cases acquired with different scanners and protocols gathered from the HaN-Seg challenge. All models were implemented using TensorFlow (version 2.1) and trained on a double NVIDIA GeForce RTX 2080Ti GPU card. Due to limitations in computational resources, particularly in GPU RAM size, the 3D U-Net architecture was trained from scratch using an isotropic patch size of 64^3 with a sliding window on its neighbouring patches with an overlap of 32^3 using the Patchify library[9]. This overlap ensures that a continuous whole-label output can be obtained and allows for an increased training dataset for the network (Fu *et al* 2020). Four experimental scenarios were formulated to determine how different tuned parameters under limited training datasets can impact the final predicted bone classes. The details of each experimental case are summarized below:

1. **Experiment I:** training network with all extracted patches and without augmentation.
2. **Experiment II:** training network with reduced background patches and without augmentation.
3. **Experiment III:** training network with reduced background patches and with augmentation.
4. **Experiment IV:** training network with reduced background patches, class-dependent augmentation and weight maps.

Aside from the first experimental training, the number of background patches (i.e. patches which do not contain any bone information) was drastically reduced by randomly selecting a few of these patches as part of the training samples. The purpose of this step is to reduce the high-class imbalance skew towards background patches. The image augmentation adopted reflect standard strategies specialized to radiotherapy application such as rotations ($\pm 15°$), shifts, flips and various amounts of image noise were applied to the training patches only.

[9] https://pypi.org/project/patchify/.

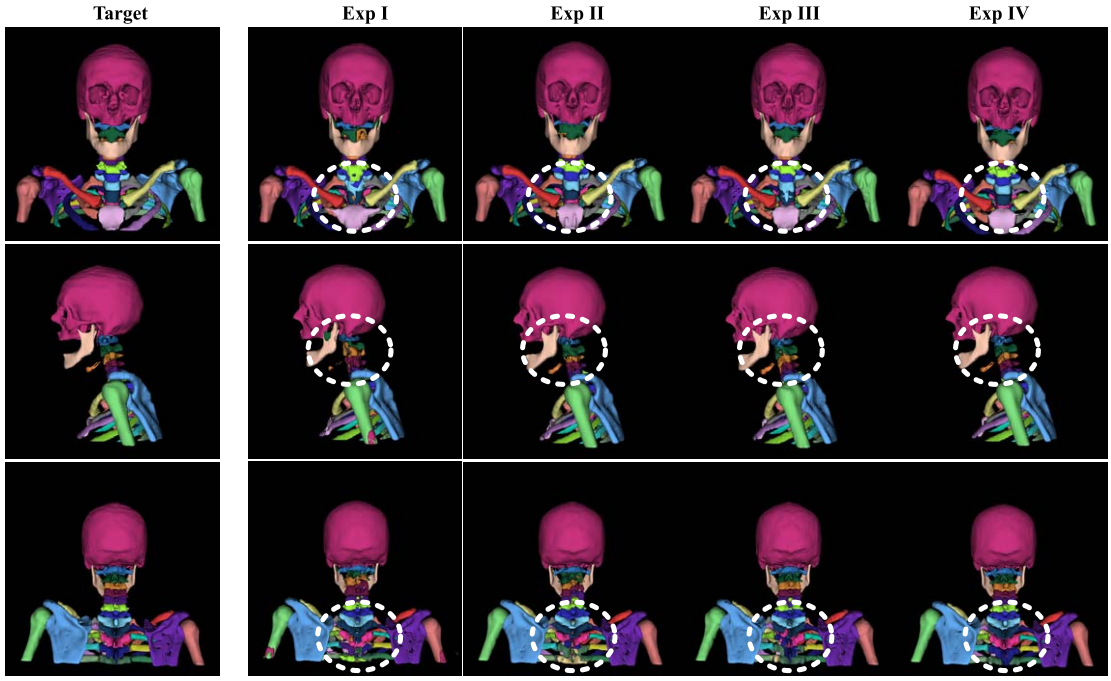

Figure 4. 3D volume rendering of target labels against its predicted labels for all experimental cases as defined in subsection 2.5 in the front, side and back view (top to bottom) of an exemplary test case in **Test A**. Different colours represent the different individual bones as explained in figure 2. **Target:** ground-truth individual bones; **Exp I:** predicted bone labels from Experiment I; **Exp II:** predicted bone labels from Experiment II; **Exp III:** predicted bone labels from Experiment III and **Exp IV:** predicted bone labels from Experiment IV. The white dashed circles on the predicted bone labels show the focus regions where improvement was observed across the different experimental cases.

The class-dependent augmentation adopted in the last experimental case addresses the class imbalance problem and it involves augmenting patches with vertebrae or ribs at a higher degree than patches with other classes using a ratio of 2:1 respectively. Besides the class-dependent augmentation in the final experimental scenario, a weight map was introduced in the loss function expression in order to compensate for the remaining class imbalance. The weight of each class was generated using equation (1) inspired by the work from La Rosa (2017)

$$W_c = \frac{\text{freq}_m}{\text{freq}_c}, \tag{1}$$

where W_c is the weight of the given bone class, freq_m is the median of the frequencies of all bone classes and freq_c is the frequency of the given bone class.

The adaptive moment estimation (Adam) (Singarimbun *et al* 2019) was the optimizing algorithm adopted with a learning rate of 2e-4. The batch size was limited to 4 with a fixed epoch of 25. These hyper-parameters were tuned empirically. To quantify the deviation of the estimated label from the target label, a combination loss of Dice and cross-entropy loss was minimized as the objective function. Once a model was trained, the model was tested using the held-out test datasets. From the predicted patches, a whole 3D volume label was obtained via patch fusion of the predicted patches using the Patchify library[10].

2.6. Post-processing

Default labels were recovered from the raw prediction of grouped labels using connected components analysis (Silversmith 2021) to remove redundant predictions in the background as well as separating grouped labels to left and right instances of bone classes such as the ribs, clavicles, humeri and scapulae. This algorithm can successfully distinguish between the left and right sides of the same bone, since the left and right sides of the same bone are not connected. Once the left and right instances have been differentiated, the labels were renamed to the same numeric value as the default labels.

2.7. Benchmark segmentation tools

The benchmark segmentation tools utilized for comparison are the nnU-Net and Totalsegmentator V1 (Wasserthal *et al* 2022). nnU-Net employs self-configuring preprocessing, augmentation and post-processing

[10] https://pypi.org/project/patchify/.

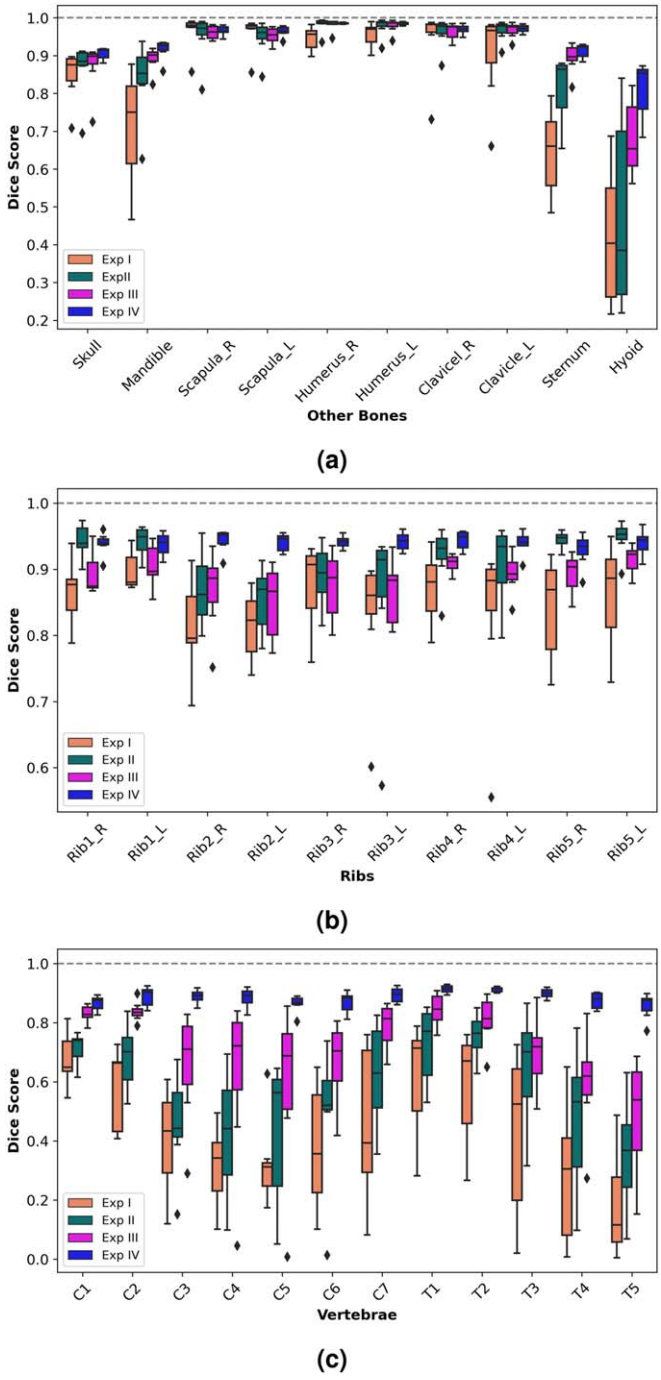

Figure 5. Boxplot analysis (Dice score) of all experimental cases as defined in subsection 2.5 using **Test A**. (a) All individual bones in the head and neck region except ribs and vertebrae (Other bones); (b) all ribs and (c) All vertebrae. Exp I—Experiment I, Exp II—Experiment II, Exp III—Experiment III and Exp IV—Experiment IV.

techniques while Totalsegmentator is a segmentation tool trained on 1228 CT images to classify over 100 classes of the human anatomy using nnU-Net. Bone predictions from Totalsegmentator were acquired using the 3D Slicer (Pieper *et al* 2004) extension (version 5.2.0). The retrained nnU-Net on our dataset was implemented using the default configuration of nnU-Net.

3. Results

The accuracy of the predicted bone labels and its corresponding target bone labels were evaluated both qualitatively and quantitatively. Qualitative and quantitative analysis was performed using visual inspection and the Dice similarity index (Thada and Jaglan 2013) respectively.

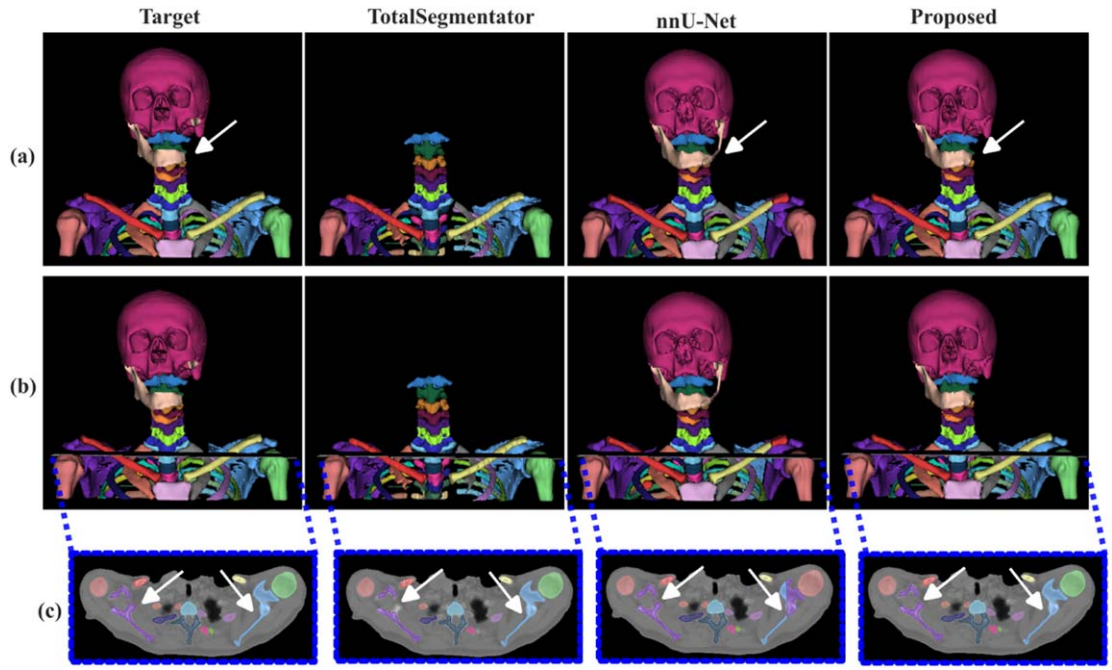

Figure 6. Comparison of our final experimental results (Experiment IV) with Totalsegmentator and a retrained nnU-Net for one test case in **Test A**. Row (a) represents the 3D volume rendering of target bone labels against predicted bone labels from Totalsegmentator, nnU-Net and our proposed approach respectively; Row (b) represents the 2D slice extraction from all cases and Row (c) represents the 2D image slice of bone labels overlaid on the planning CT for all cases. The white arrow points to the region where large deviations can be observed for the different approaches.

3.1. Comparison between experimental scenarios

Test A representing the cohorts of DKFZ, HIT, TCIA and UKHD as reflected in the training datasets was utilized to analyze the different experimental scenarios. Figure 4 outlines the 3D volume rendering of target bone labels against its corresponding predicted bone labels in the front, side and back view on a sample test case for all experimental scenarios. Quantitative results from the dice index measure were classified into separate boxplots of ribs, vertebrae and other bones consisting of skull, mandible, hyoid, sternum, clavicles, humeri and scapulae, as detailed in figure 5. The trends observed in the grouped boxplots were in agreement with the visual inspection. Each experimental case resulted in a unique prediction with significant differences observed in various bone types. In view of the first experimental scenario, large deviations were observed in all individual bones and this is due to the large number of background patches present in the training datasets. These patches hinder effective learning although it provides huge dataset for the network even without any form of augmentation. Therefore, background patches were drastically reduced as described in Experiment II. Despite the few patches in Experiment II, this step enables the network to focus on patches which have useful information to effectively learn and resulted in a more precise prediction than Experiment I. This parameter adaptation led to a significant improvement in large bones such as the skull, mandible, scapulae, humeri and clavicles. Nevertheless, large variations observed in bone classes such as ribs and vertebrae persisted in this experimental case. This can be attributed to the less representation of these bone classes within the training sample. Experiment III comprising data augmentation yielded an enhancement in all individual bones including ribs and vertebrae. However, the mixing of vertebral bones with its neighbouring vertebrae was still present. Thus, with a focus on vertebral bone, additional steps were taken to further improve the prediction accuracy of these bone classes as demonstrated in Experiment IV. Experiment IV produced more comparable vertebral bones to their target with a reduced mix-up of neighbouring vertebrae.

3.2. Comparison with benchmark segmentation approaches

The comparison of our final experimental results (Experiment IV) with Totalsegmentator and the retrained nnU-Net is depicted in figure 6. Predictions from Totalsegmentator was unable to predict bone classes such as the skull, mandible, hyoid and sternum. Additionally, ribs were segmented without their corresponding costal cartilages. Unlike the Totalsegmentator, the retrained nnU-Net on our datasets was able to predict all individual bones in the head and neck region. The retrained nnU-Net predictions yielded comparable results with our proposed approach, and large deviations were mostly observed in the left and right instances of the same bone classes, such as the ribs, clavicles, humeri and scapulae, as reflected in the boxplot analysis in figure 7.

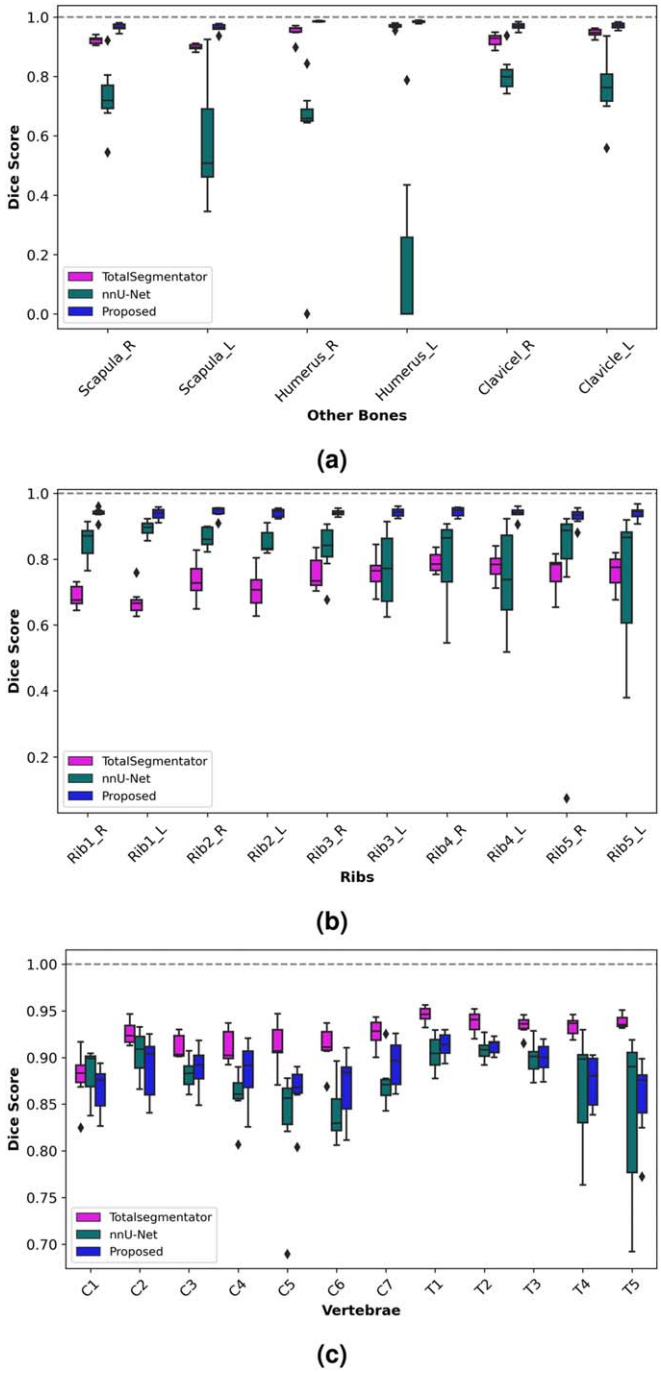

Figure 7. Boxplot analysis (Dice score) of our proposed solution approach (Experiment IV) with Totalsegmentator and nnU-Net using **Test A**. (a) All individual bones in the head and neck region except ribs and vertebrae (Other bones); (b) all ribs and (c) all vertebrae.

3.3. Evaluation on a new cohort

Test B was used to evaluate the generalization capability of our proposed model on patients acquired with different scanners and acquisition parameters. As demonstrated in figure 8, the prediction accuracy was on a similar level as **Test A** for most individual bones with fluctuations observed in thoracic vertebrae (especially T4 and T5) between different patients. That is, our model accurately identified the different thoracic vertebrae without mixing neighbouring thoracic vertebrae for some patient cases and vice versa. A detailed quantitative analysis is illustrated in figure 9.

4. Discussion

In this study, we provide insights on essential parameters needed to automatically and accurately segment all individual bones in the head and neck region using very limited planning CT images of patients who received

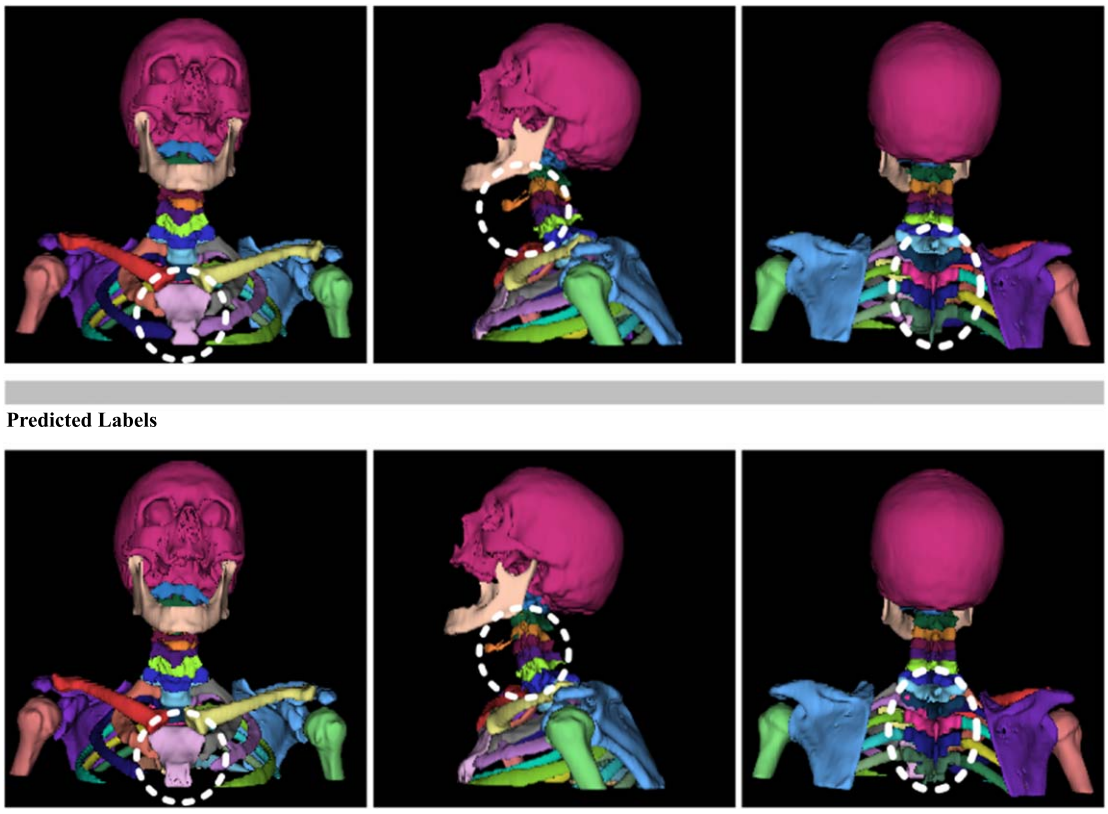

Figure 8. 3D volume rendering of target bone labels against its predicted bone labels for one case in **Test B**. Different colours represent the different individual bones as shown in figure 2. The white dashed circles show regions where deviations are visible.

radiotherapy. Before our study, preceding published works which adopted similar network architectures to model different scales of input CT (diagnostic CT or planning CT) as well as bone classes have proven that further validation and evaluation are needed for stand-alone deep learning based approaches before their introduction into clinical routines.

The performance of deep learning models grow with the amount of data in the training dataset (Sarker 2021). However, the performance of individual deep learning models can be saturated after the addition of certain amount of data (Shorten and Khoshgoftaar 2019). Thus, further improvements can frequently be attained by either extending model architectures or exploring parameters within the available datasets which will enforce effective learning from limited datasets. For this reason, we have examined 4 experimental cases with the aim of determining how different parameters can impact the final predicted bone classes on 30 patient studies. These parameters include background patch reduction, class-dependent augmentation and the incorporation of weight information to the loss function of the network model parameters. From all the different sets of experiment carried out, we can deduce that the sequential tuning of different parameters provided the network with rich information during the training phase to help differentiate all individual bones present in the head and neck region irrespective of their shape or size.

In figures 6 and 7, we compared our proposed approach (Experiment IV) with Totalsegmentator and a retrained nnU-Net on datasets. As mentioned earlier, Totalsegmentator has been trained on a large number of CT images (both diagnostic and planning). Therefore, the high prediction accuracy observed in all vertebral bones (figure 7(c)) by Totalsegmentator is expected; because the model has seen many representations of vertebral bones. Nonetheless, it was unable to segment some bone classes (skull, mandible, hyoid and sternum) and ribs were segmented without their corresponding costal cartilages. Since our research question is focused to only radiotherapy (i.e. planning CT) and learning from limited datasets, the nnU-Net was retrained on datasets to promote fair comparison. Predictions from nnU-Net resulted in an accurate prediction in the upper vertebrae (such as C1 and C2) with large variance observed in the lower vertebrae for all test cases. Even so, as reported in the findings of Schnider *et al* (2020), nnU-Net retrained on our datasets was unable to distinguish between the left and right instances of the same bone classes such as the scapulae, clavicles, humeri and the mix-up of rib classes; which our model was able to solve by the adopted grouping technique prior to the model training.

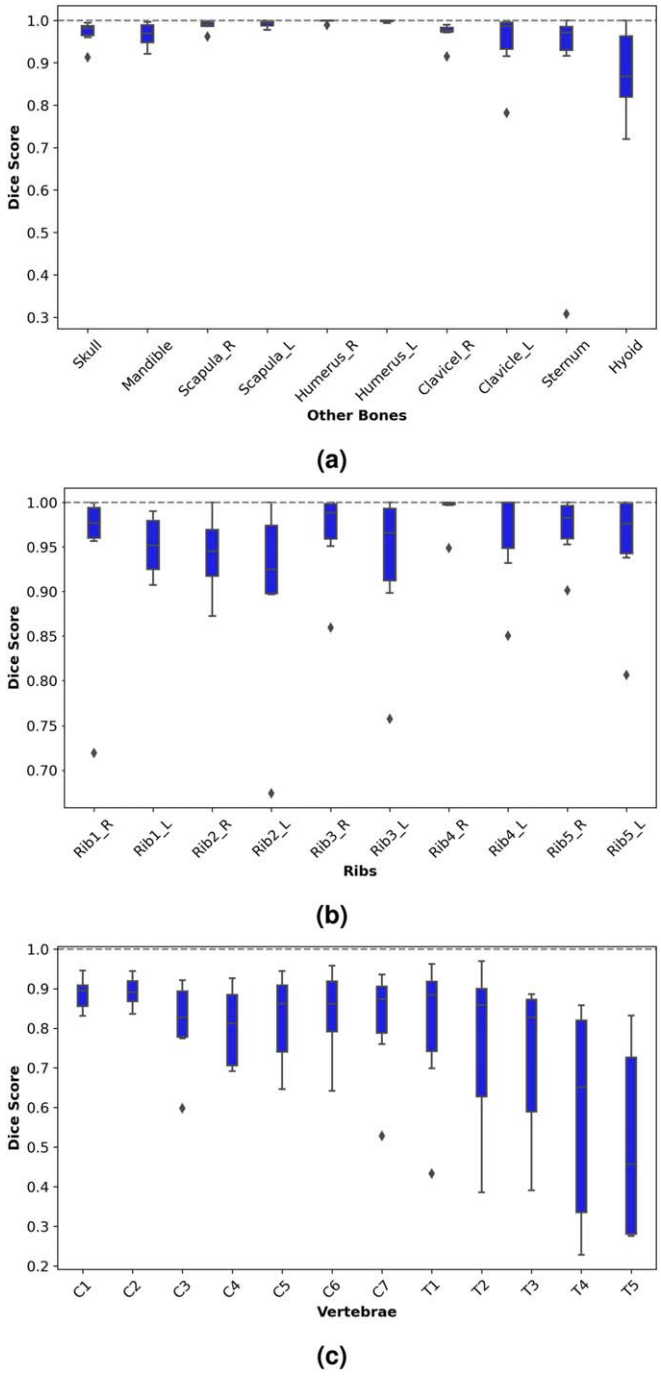

Figure 9. Boxplot analysis (Dice score) of our proposed solution approach (Experiment IV) using **Test B**. (a) All individual bones in the head and neck region except ribs and vertebrae (Other bones); (b) all ribs and (c) all vertebrae.

Furthermore, to quantify how our model generalizes on a new cohort, we tested our model on 8 additional test cases from the ongoing HaN-Seg challenge. The data selection criteria was based on their similarity in terms of matrix dimension and spatial resolution. The detailed analysis as demonstrated in figure 9 proves an optimal generalization to new patient cases in almost all individual bones. Nevertheless, large fluctuations observed in the prediction accuracy of thoracic vertebrae (especially T4 and T5) can be attributed to the different patient positioning set-up during the delivery of radiotherapy. Patient positioning has a great impact on the appearance of thoracic spine (Plataniotis *et al* 2019) and it confused our model by mixing up neighbouring thoracic vertebrae since our selected U-Net architecture is variant to rotation. Therefore, as a future direction, we will further extend our model by investigating and introducing a rotational component in the class-dependent augmentation parameter to account for the change in the appearance of thoracic vertebrae due to the different patient positioning.

5. Conclusion

This study examined essential parameters needed to automatically and accurately segment individual bones on planning CT images of head and neck cancer patients. We explored 4 experimental scenarios to determine the impact of different parameter adaptations on predicted individual bones irrespective of their shape, size and level of complexity on 30 patients. Our proposed solution approach proves that the sequential tuning of parameters such as background patch reduction, class-dependent augmentation and the incorporation of weight information in the loss function yielded an equivalent prediction with a retrained nnU-Net on our datasets for all vertebral bones and outperformed nnU-Net in the correct identification of left and right instances of the same bone class; because of the grouping approach adopted before the model generation. Furthermore, our model showed an optimal generalization capability for almost all individual bones on a new test cohort. Nevertheless, large fluctuations observed in thoracic vertebrae warrant the introduction of a rotational component in the class-dependent augmentation parameter to account for the different patient positioning. With these insights, we are challenging the utilization of an automatic and accurate bone segmentation tool into the clinical routine of radiotherapy despite the limited training datasets.

Acknowledgments

Ama Katseena Yawson, **Kristina Giske** and **Sebastian Klüter** are funded by the German Federal Ministry of Education and Research within the track 'Bildgeführte Diagnostik und Therapie—Neue Wege in der Intervention' in ARTEMIS project (13GW0436). **Alexander Walter** is funded by the Helmholtz Information and Data Science School for Health (HIDSS4Health).

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Ethical statement

This retrospective study was reviewed and approved by the Institutional Review Board of University Clinic Heidelberg (S-660/2022). The study was performed in accordance with the principles embodied in the Declaration of Helsinki and in accordance with the professional regulations of the State Medical Association of Baden-Württemberg.

ORCID iDs

Ama Katseena Yawson https://orcid.org/0000-0002-6005-7354
Alexandra Walter https://orcid.org/0000-0001-5044-0133
Nora Wolf https://orcid.org/0009-0001-3789-4234
Sebastian Klüter https://orcid.org/0000-0003-3139-3444
Oliver Jäkel https://orcid.org/0000-0002-6056-9747
Kristina Giske https://orcid.org/0000-0001-7287-3271

References

Ang K K *et al* 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** 2940

Balagopal A, Kazemifar S, Nguyen D, Lin M-H, Hannan R, Owrangi A and Jiang S 2018 Fully automated organ segmentation in male pelvic CT images *Phys. Med. Biol.* **63** 245015

Bauer C J, Teske H, Walter A, Hoegen P, Adeberg S, Debus J, Jäkel O and Giske K 2023 Biofidelic image registration for head and neck region utilizing an in-silico articulated skeleton as a transformation model *Phys. Med. Biol.* **68** 095006

Belal S L, Sadik M, Kaboteh R, Enqvist O, Ulén J, Poulsen M H, Simonsen J, Høilund-Carlsen P F, Edenbrandt L and TrägÅrdh E 2019 Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal metastases *Eur. J. Radiol.* **113** 89–95

Bosch W R, Straube W L, Matthews J W and Purdy J A 2015 Data from head-neck_cetuximab *Cancer Imaging Archive* **10** K9

Çiçek Ö, Abdulkadir A, Lienkamp S S, Brox T and Ronneberger O 2016 3D u-net: learning dense volumetric segmentation from sparse annotation *Int. Conf. on Medical Image Computing and Computer-assisted Intervention* (Springer) pp 424–32

Fu Y, Lei Y, Wang T, Tian S, Patel P, Jani A B, Curran W J, Liu T and Yang X 2020 Pelvic multi-organ segmentation on cone-beam ct for prostate adaptive radiotherapy *Med. Phys.* **47** 3415–22
Giske K *et al* 2011 Local setup errors in image-guided radiotherapy for head and neck cancer patients immobilized with a custom-made device *Int. J. Radiat. Oncol.* Biol.* Phys.* **80** 582–9
Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Int. Conf. on Machine Learning, PMLR* pp 448–56
Isensee F, Jaeger P F, Kohl S A, Petersen J and Maier-Hein K H 2021 nnu-net: a self-configuring method for deep learning-based biomedical image segmentation *Nat. Methods* **18** 203–11
Karimi S, Cosman P, Wald C and Martz H 2012 Segmentation of artifacts and anatomy in CT metal artifact reduction *Med. Phys.* **39** 5857–68
Klein A, Warszawski J, Hillengaß J and Maier-Hein K H 2019 Automatic bone segmentation in whole-body CT images *Int. J. Comput. Assist. Radiol. Surg.* **14** 21–9
Kompella G, Antico M, Sasazawa F, Jeevakala S, Ram K, Fontanarosa D, Pandey A K and Sivaprakasam M 2019 Segmentation of femoral cartilage from knee ultrasound images using mask r-CNN *2019 41st Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) (Berlin, Germany, 23-27 July 2019)* (IEEE) pp 966–9
La Rosa F 2017 *A Deep Learning Approach to Bone Segmentation in CT scans* Universit`a di Bologna
Leydon P, O'Connell M, Greene D and Curran K M 2022 Bone segmentation in contrast enhanced whole-body computed tomography *Biomedical Physics & Engineering Express* **8** 5
Meyer E C, Walser R, Hermann B, La Bash H, DeBeer B B, Morissette S B, Kimbrel N A, Kwok O-M, Batten S V and Schnurr P P 2018 Acceptance and commitment therapy for co-occurring posttraumatic stress disorder and alcohol use disorders in veterans: pilot treatment outcomes *J. Traumatic Stress* **31** 781–9
Minnema J, van Eijnatten M, Kouw W, Diblen F, Mendrik A and Wolff J 2018 CT image segmentation of bone for medical additive manufacturing using a convolutional neural network *Comput. Biol. Med.* **103** 130–9
Möller T B 2005 *Taschenatlas der Schnittbildanatomie: Kopf, Hals* (Georg Thieme Verlag) vol 1
Perona P, Shiota T and Malik J 1994 Anisotropic diffusion *Geometry-driven Diffusion in Computer Vision* (Springer) pp 73–92
Pieper S, Halle M and Kikinis R 2004 3D slicer *2004 2nd IEEE Int. Symp. on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821), IEEE* pp 632–5
Plataniotis N, Evangelopoulos D, Katzouraki G and Pneumaticos S 2019 The effect of patient positioning on the relative position of the aorta to the thoracic spine *Eur. Spine. J.* **28** 477–83
Podobnik G, Strojan P, Peterlin P, Ibragimov B and Vrtovec T 2023 HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset *Medical physics* **50** 1917–27
Sahiner B, Pezeshk A, Hadjiiski L M, Wang X, Drukker K, Cha K H, Summers R M and Giger M L 2019 Deep learning in medical imaging and radiation therapy *Med. Phys.* **46** e1–36
Sarker I H 2021 Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions *SN Comput. Sci.* **2** 420
Schnider E, Horváth A, Rauter G, Zam A, Müller-Gerbl M and Cattin P C 2020 3d segmentation networks for excessive numbers of classes: distinct bone segmentation in upper bodies *Int. Workshop on Machine Learning in Medical Imaging* (Springer) pp 40–9
Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning *J. Big Data* **6** 1–48
Silversmith W (2021) *cc3d: Connected Components on Multilabel 3D Images* (https://doi.org/10.5281/zenodo.5719536)
Singarimbun R N, Nababan E B and Sitompul O S 2019 Adaptive moment estimation to minimize square error in backpropagation algorithm *2019 Int. Conf. of Computer Science and Information Technology (ICoSNIKOM)* (IEEE) pp 1–7
Stoiber E M, Lechsel G, Giske K, Muenter M W, Hoess A, Bendl R, Debus J, Huber P E and Thieke C 2009 Quantitative assessment of image-guided radiotherapy for paraspinal tumors *Int. J. Radiat. Oncol.* Biol.* Phys.* **75** 933–40
Stoll M, Stoiber E M, Grimm S, Debus J, Bendl R and Giske K 2016 Comparison of safety margin generation concepts in image guided radiotherapy to account for daily head and neck pose variations *PLoS One* **11** e0168916
Suzuki K 2017 Overview of deep learning in medical imaging *Radiol. Phys. Technol.* **10** 257–73
Thada V and Jaglan V 2013 Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm *Int. J. Innov. Eng. Technol.* **2** 202–5
Vogelius I R, Petersen J and Bentzen S M 2020 Harnessing data science to advance radiation oncology *Mol. Oncol.* **14** 1514–28
Wahid K A *et al* 2023 Large scale crowdsourced radiotherapy segmentations across a variety of cancer anatomic sites *Sci. Data* **10** 161
Wasserthal J, Meyer M, Breit H-C, Cyriac J, Yang S and Segeroth M 2023 Totalsegmentator: robust segmentation of 104 anatomical structures in CT images *Radiology: Artificial Intelligence* **5** 5
Xu J, Li Z, Du B, Zhang M and Liu J 2020 Reluplex made more practical: leaky relu *2020 IEEE Symp. on Computers and Communications (ISCC)* (IEEE) pp 1–7
Xu L, Tetteh G, Lipkova J, Zhao Y, Li H, Christ P, Piraud M, Buck A, Shi K and Menze B H 2018 Automated whole-body bone lesion detection for multiple myeloma on 68 Ga-pentixafor PET/CT imaging using deep learning methods *Contrast Media Mol. Imaging* **2018** 11