EG UK Computer Graphics & Visual Computing (2018)
G. Tam and F. Vidal (Editors)

Segmenting Teeth from Volumetric CT Data with a Hierarchical CNN-based Approach

P. Macho, N. Kurz, A. Ulges, R. Brylka, T. Gietzen, U. Schwanecke

RheinMain University of Applied Sciences, Wiesbaden, Germany

Abstract

This paper addresses the automatic segmentation of teeth in volumetric Computed Tomography (CT) scans of the human skull. Our approach is based on a convolutional neural network employing 3D volumetric convolutions. To tackle data scale issues, we apply a hierarchical coarse-to fine approach combining two CNNs, one for low-resolution detection and one for high-resolution refinement. In quantitative experiments on 40 CT scans with manually acquired ground truth, we demonstrate that our approach displays remarkable robustness across different patients and device vendors. Furthermore, our hierarchical extension outperforms a single-scale segmentation, and network size can be reduced compared to previous architectures without loss of accuracy.

CCS Concepts

•***Computer Graphics*** → *Image processing;* •***Computing / Technology Policy*** → *Medical technologies;* •***Machine Learning*** → *Neural networks;*

1. Introduction

Recent progress in Deep Learning has given us vastly improved models for various image analysis tasks such as categorization, object detection or the estimation of scene structure and motion. Particularly, Convolutional Neural Networks (CNNs) – which learn stacked image filters tailored to the task and data at hand – have successfully been used for semantic segmentation in various domains such as street scenes [COR*16] or medical imagery [OPT15].
There are many 2D approaches [OPT15, ASM17, MMH*17, LBBH98, KSH12, NHH15, BKC17, MSH*17, SLD17], as well as approaches directly segmenting volumes in 3D [MNA16, ÇAL*16, CSA00, GFJ16, LDS*17, ZKZ*18]. This is of particular interest for volumetric imaging in the medical domain such as Computed Tomography (CT), which is commonly used for a wide range of tasks. The challenge addressed in this paper is to segment teeth in a CT volume as shown in Figure 1. We present a CNN-based model that applies 3D convolutions, following a commonly used bottleneck architecture with skip connections [MNA16] that has been applied to volumetric prostate scans before [LTvdV*14]. Our contributions are:

- Our study is – to the best of our knowledge – the first one on CNNs for a 3D volumetric segmentation of teeth. Our model is an extension of V-Net [MNA16] that addresses scale problems by a simple coarse-to-fine hierarchical extension, which first roughly localizes and then refines the teeth region.
- We present quantitative experiments on a set of 40 CT scans with different patients and devices, which show that our model displays strong robustness for uncalibrated devices and yields strong improvements over a thresholding baseline. We also investigate the influence of network size, data scaling and training data volume.

2. Related Work

CNNs: Despite early work in the 1990s [LBBH98], CNNs have become popular quite recently with AlexNet's outstanding object category recognition performance [KSH12] in the 2012 ImageNet Challenge [RDS*15]. Since SegNet [BHC15], CNNs have also been studied intensely for semantic image segmentation [COR*16].

Medical Imaging: CNNs have been applied in medical imaging for the classification of teeth based on CT slices [MMH*17] or for supervised 2D segmentation. Here, U-Net [OPT15] demonstrates the benefit of data augmentation for small datasets. Sekuboyina et al. [ASM17] use deep networks for a localization and segmentation of lumbar vertebrae in CT-scans, which is treated as a non-linear regression problem to determine bounding boxes in the volume. The corresponding multi-class segmentation is done by a modified 2D U-Net [OPT15] trained on sagittal slices. In [MRAG08] a segmentation of teeth in CT data is performed by using panoramic resampling of 2D coronal slices and variational level set to determine the teeth contour.

CNNs for Volumetric Data: Recently, first deep learning models have been demonstrated to deal with 3D data directly. Advan-

DOI: 10.2312/cgvc.20181213

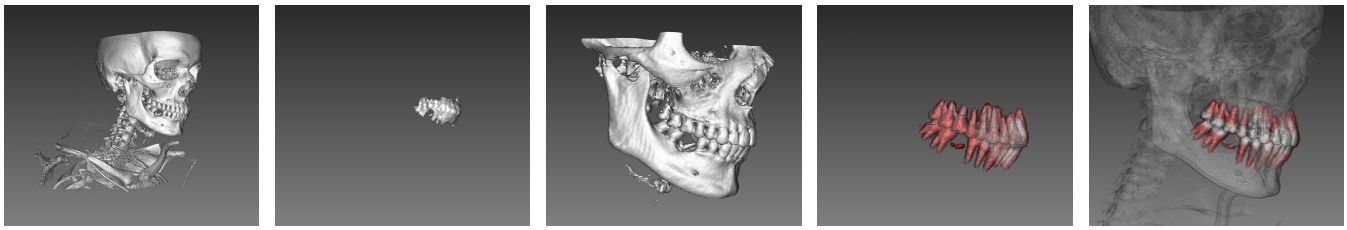

Figure 1: *The overall workflow of our model (numbered left to right): The input volume is (1) rescaled to* 128^3*, processed by the low-res model and refined, obtaining an ROI* $\mathcal{R}$ *(2). An isotropic* 128^3 *high-resolution volume containing* $\mathcal{R}$ *(3) is then segmented by the high-res model, obtaining a fine-grain segmentation. Pictures (4)+(5) show the result (red) overlayed with ground truth / transparent input volume. A dental root missed in the ground truth is detected.*

tages of this approach (as opposed to stacking 2D slices) are emphasized in the survey by Ioannidou et al. [ICNK17]. For example, VoxNet [MS15] uses a 3D CNN for real time object recognition using three different 3D data sources (LiDAR point clouds, RGBD point clouds, and CAD models) and achieved state-of-the-art accuracy. Closest to our work is Milletari et al.'s V-Net model [MNA16], which applies a CNN approach in the domain of 3D medical image segmentation and has proven successful in segmenting prostates within MRI scans.

3. Approach

Following internal ethical review board approval, head CT scans were collected from the PACS system of the University Medical Center Mainz. We only used existing CT data (from four different CT devices) from the DICOM database.

No subject was exposed to ionizing radiation for this research. The local ethical approval board [Eth] has approved the processing of the pseudonymized existing CTs.

Our model processes volumetric inputs of size $512 \times 512 \times 512$ (or 512^3). In this context, scale poses a challenge: While 128^3 inputs were found to fit an 8 GB GPU memory, the full 512^3 volume cannot be processed with standard GPUs simultaneously. Therefore, we choose a hierarchical approach similar to Sekuboyina et al.'s [ASM17], in which a coarse localization of a *region of interest* (ROI) within the downscaled volume is followed by a fine-grain segmentation at full resolution. Unlike Sekuboyina et al., who apply different approaches for both steps, we use two similar 3D-CNN segmenters trained on different resolutions: The first (called the *low-res model*) localizes the region of interest, of which the second (the *high-res model*) produces a fine-grain segmentation. By combining both networks, our model yields a fine-grain segmentation of the full 512^3 input volume. The workflow of our model is also illustrated in Figure 1.

Base Model: Both low-res and high-res model are 3D-CNN segmenters, i.e. they process volumetric input data and output voxel-wise posteriors, which we threshold to obtain binary segmentation masks. Both models are based on V-Net [MNA16], a fully convolutional 3D-CNN that combines a bottleneck architecture with skip connections, where bottom layers compress the input by applying stacked convolutions, and top layers use de-convolutions to combine the compressed data with the extracted feature maps from earlier layers to preserve fine details. This results in voxel-wise class labels. The network's parameters (convolution and de-convolution maps) are trained on labeled volumes using backpropagation. We increased the input shape from the original $128 \times 128 \times 64$ to 128^3 and investigated the influence of channel reduction (see Section 4). Our implementation is based on the deep learning framework Caffe [JSD*14] with Milletari's 3D extension [Fau].

Low-res Model and ROI Localization: The low-res model aims at roughly localizing the dental region. To do so, we downsample the input volume to 128^3 voxels by trilinear interpolation, such that the network can process the volume at once. The result is a 3D segmentation map containing voxel-wise posteriors between 0% and 100%. By thresholding at $t_1 = 50\%$, we obtain a foreground (tooth) region of interest $\mathcal{R}$ (see Figure 1, second plot). We refine $\mathcal{R}$ by applying connected component labeling and dropping all components except the largest, which removes spurious false positives. The refined $\mathcal{R}$ serves as the input for further processing.

ROI Normalization and High-res Model: The second model takes the ROI $\mathcal{R}$ as input and produces a fine-grain segmentation. Note that – like for the low-res model – input data of 128^3 voxels are required. Also, the real-world voxel size is anisotropic and differs vastly between CT devices and vendors (see Table 1). Therefore, the input volume is resampled to high-resolution *isotropic* voxels. Afterwards, we select a 128^3 subregion containing the tooth region. If $\mathcal{R}$ is too big to fit 128^3 voxels (e.g., due to connected artifacts), we increase the probability threshold t_1 beyond 50% to reduce $\mathcal{R}$ until it fits. From this input, the high-res model produces a 3D foreground posterior map. Finally we apply a second threshold $t_2 = 50\%$ to obtain the final segmentation.

Training Procedure: Both models are trained from scratch on 40 CT volumes by maximizing the dice loss

$$\mathcal{L} = \frac{2\sum_i p_i r_i}{\sum_i p_i + \sum_i r_i}$$

between ground truth r_i and voxel-wise posteriors p_i using SGD (*Stochastic Gradient Decent*). For the low-res model, we use data augmentation – namely histogram matching across different volumes, deformation and translation in x, y, z direction – to increase the number of training samples. Additionally, we apply linear standardization to all volumes' voxel values, with means and standard deviations estimated over the whole set.
For the high-res model, we apply data augmentation by cropping

randomly shifted patches containing the ground truth ROI out of a volume which is resampled to isotropic spacing. We trained both models for around 36,000 iterations, where training was found empirically to have converged. Just like for the low-res model, we used linear standardization.

4. Experiments

We systematically studied the impact of reducing the number of channels on the segmentation results. We also compared our results to a threshold method that serves as a baseline. Our experiments are based on 40 CT cephalic samples captured by four different devices from two different manufacturers. All data have size 512^3 but differ in spatial dimensions, spacing and dynamic range. Table 1 gives an overview of the properties of our dataset.

Device	#	Spacing [x,y,z]
Toshiba Aquilion	20	[0.41–0.64, 0.41–0.64, 0.30–0.30]
Philips iCT 256	17	[0.37–0.82, 0.37–0.82, 0.50–0.65]
Philips Brilliance 64	2	[0.60–0.68, 0.60–0.68, 0.80–0.80]
Philips Mx8000 IDT 16	1	[0.57–0.57, 0.57–0.57, 1.00–1.00]

Table 1: *Characteristics of our dataset*

To acquire ground truth we implemented an application especially designed to support the manual labeling of teeth in volumetric data. Our implementation first determines an individual threshold for each dataset used to separate soft tissue and artifacts from bony structures and teeth. Since some artifacts persist and bones and teeth cannot be perfectly separated, the remaining data is labeled manually.

Training Data Size: To ensure that our approach does not depend on the characteristics of a particular vendor, we trained our network several times with different training and validation sets. Each time, the data are split into a training set using 90% and a validation set using 10% of the samples, each set containing samples from every manufacturer. After approximately 12,000 training steps, we achieved a dice loss between 0.83 and 0.87 on the validation data. This shows that the proposed approach is working even across various manufacturers.

Network Size: Volumetric data is far more complex than 2D image data. Thus, a neural network that works with volumes usually is much larger and therefore requires much more memory compared to a network that works with images. As memory (especially GPU memory) is a limited resource, this can be problematic. To cope with this problem we reduced the size of the network by reducing the number channels as much as possible without reducing the accuracy on the data. A beneficial side effect of a smaller network is a reduction in overall training time, since the network has fewer parameters that need training.

We decreased the size of the network relative to the original V-Net [MNA16] by reducing the number of channels in each inner layer by the same fixed factor, while the outer layers kept their original shape. This was done on both the low-res model and the high-res model with similar results. Figure 2 shows the dice loss during the training with a reduced number of channels on the high-res model. As can be seen, the network performs quite well for all reduction factors without significantly affecting accuracy on the data. Only the reduction by a factor of 8 appears to be problematic. These results are also confirmed by Table 2 which shows the final loss on the low-res model. The table also shows that a smaller network can lead to a higher dice loss on our data, while reducing the time needed per training iteration. The best results were achieved by a channel reduction by a factor of 2 and 4, which leads to a high dice loss while simultaneously reducing training time and memory consumption considerably.

#Channels div by	Final dice loss	Time	Memory Usage
1	0.7986	12 s	7987 MiB
2	0.8017	10 s	6436 MiB
4	0.7849	9 s	5792 MiB
8	0.7689	9 s	5504 MiB
16	0.0183	9 s	5368 MiB

Table 2: *The impact of the channel reduction on loss, training time per iteration, and total memory usage of the low-res model.*

Comparison with Baseline: The segmentation of bones or soft tissue in CT data is often done via simple thresholding. Especially with teeth this poses a challenge, since the density of teeth is close to or even equal to the density of bones. In addition, it is problematic to compare data from different (sometimes uncalibrated) devices with different gray scale quantizations and resolutions (see e.g. [PJSM15]).

We compare our approach with a baseline that maximizes the dice loss of a thresholded input volume. To get this baseline, we increased the segmentation threshold from 1,000 to 5,000 Hounsfield Units (HU) with step size 1 and calculate the corresponding dice loss between ground truth and the result of the threshold-based segmentation. Figure 3 shows the worst and best case result of the threshold based segmentation for our 40 scans. The diagram on the right side of Figure 2 shows the relationship between the threshold used and the corresponding dice loss. It can be clearly seen that different CT devices lead to different optimal thresholds. Even for the same manufacturer it is not possible to define one fixed threshold to get the best results. The plot also includes the averaged results of our models (dashed lines).

Table 3 gives an overview of different approaches. We confirm again that the threshold baseline fails (43.3%). Also, our hierarchical approach gives significant improvements over a single-scale segmentation at low resolution (low-res model, 81.3%). The best result (93.4%) is obtained when applying the high-res model at the finest voxel spacing of 0.85 mm (which is supported by 38 of the 40 scans). Increasing the voxel spacing to 1 mm comes with a slight decrease in accuracy (but is supported by all scans).

Figure 4 illustrates the results of our full hierarchical model for test data unseen in training, comparing model outputs (left, red) with the corresponding ground truth (white, right). High-res segmentation was applied at 0.85^3 voxel spacing. The top row shows the worst result from our dataset with a dice loss of 88%. The second row shows an average dice loss around 92% and the last row our best result of 94%.

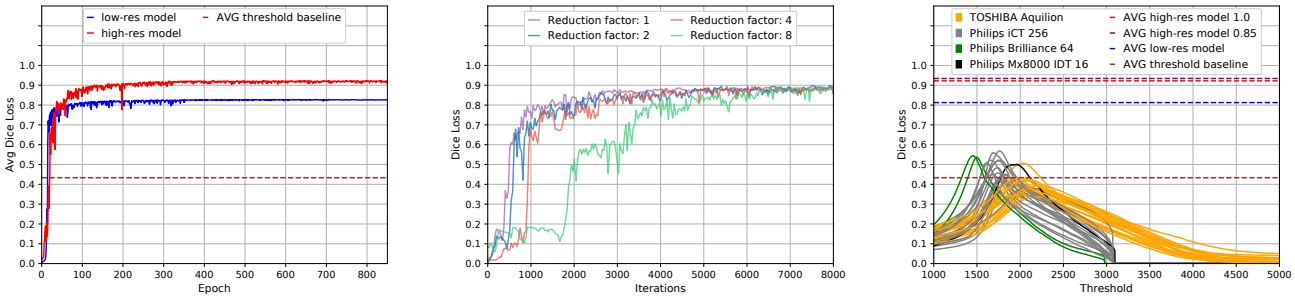

Figure 2: *Left: Averaged dice loss of the low-res model (blue) and the high-res model (red) for each epoch. Middle: Different channel reduction factors on the high-res model. Right: Result of the threshold baseline approach. Each color identifies a device. Dashed lines are averaged results of the low-res model, high-res model and threshold baseline respectively.*

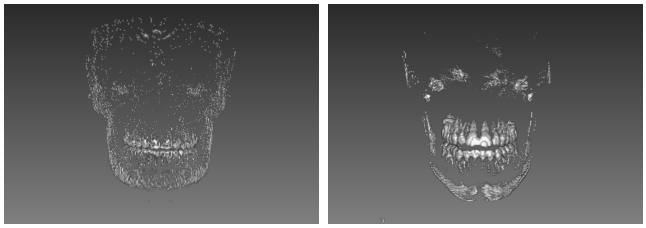

Figure 3: *Threshold baseline result. The worst result is on the left, the best result on the right.*

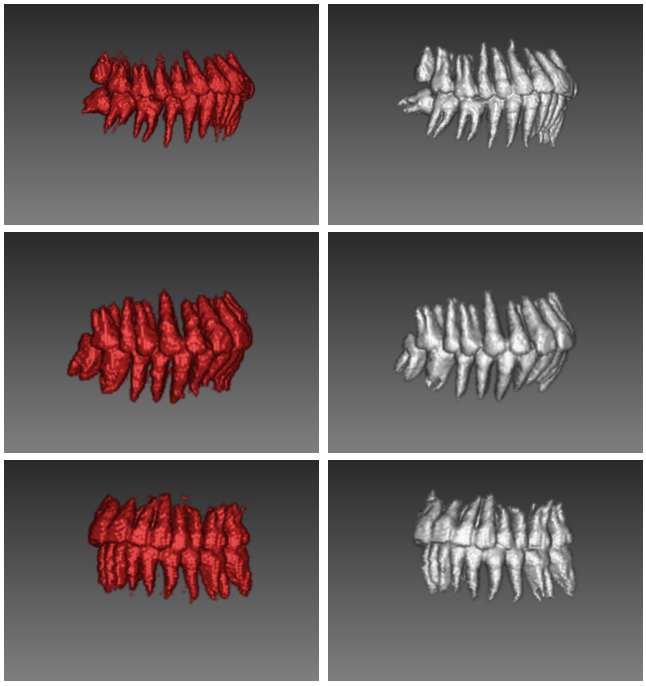

Figure 4: *Results of our full hierarchical model for scans unseen in training. The output of the model is on the left (red), ground truth on the right (white). The dice loss of the results ranges from* 88% *(top) to* 94% *(bottom).*

Approach	Spacing	avg. dice loss (%)	avg. tooth ROI
low-res model	original$\times 4$	81.26	$31 \times 26 \times 30$
hierarchical	1.0^3	92.22	$64 \times 55 \times 47$
hierarchical	0.85^3	93.42	$75 \times 65 \times 55$
threshold baseline	original	43.33	$125 \times 107 \times 123$

Table 3: *Quantitative results of different approaches. The last column shows the average size of the teeth region in relation to the used approach and resolution.*

5. Conclusions

We have applied 3D CNNs for the segmentation of teeth in CT volumes, and demonstrated that a simple hierarchical extension to V-Net [MNA16] can tackle data scale issues effectively. Our approach shows remarkable robustness: The ROI resulting from the low-res model contains the correct dental region in 100% of cases, and the high-res model yields a stable segmentation of the teeth themselves, even detecting false negatives in the ground truth (see Figure 1, Picture 4+5). According to visual verification and the dice loss measure, our approach performs a stable result comparable to the ground truth. From a broader perspective, our hierarchical coarse-to-fine approach – which trains resolution-specific CNNs for detection of an ROI and its fine-grain segmentation – can be adapted to a wide range of tasks dedicated to region based segmentation within large-scale volumes, where limited hardware is an issue.

6. Acknowledgements

This work is part of the research project “Kephalos” funded by the Federal Ministry of Education and Research. We gratefully acknowledge the Department of Diagnostic and Interventional Radiology, University Medical Center of the Johannes Gutenberg University Mainz, Germany for providing us with the DICOM-data. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

References

[ASM17] ANJANY SEKUBOYINA ALEXANDER VALENTINITSCH J. S. K., MENZE B. H.: A Localisation-Segmentation Approach for Multi-label Annotation of Lumbar Vertebrae using Deep Nets. *CoRR abs/1703.04347* (2017). 1, 2

[BHC15] BADRINARAYANAN V., HANDA A., CIPOLLA R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. *arXiv:1505.07293* (2015). 1

[BKC17] BADRINARAYANAN V., KENDALL A., CIPOLLA R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 12 (Dec 2017), 2481–2495. 1

[ÇAL*16] ÇIÇEK Ö., ABDULKADIR A., LIENKAMP S., BROX T., RONNEBERGER O.: 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2016), pp. 424–432. 1

[COR*16] CORDTS M., OMRAN M., RAMOS S., REHFELD T., ENZWEILER M., BENENSON R., FRANKE U., ROTH S., SCHIELE B.: The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR* (2016), IEEE Computer Society, pp. 3213–3223. 1

[CSA00] CARR H., SNOEYINK J., AXEN U.: Computing Contour Trees in All Dimensions. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2000), SODA '00, Society for Industrial and Applied Mathematics, pp. 918–926. 1

[Eth] Ethik-Kommission der Landesärztekammer Rheinland-Pfalz Deutschhausplatz 2, 55116 Mainz. Approval number: No 837.244.15 (10012)(05.08.2015). 2

[Fau] FAUSTO MILLETARI: 3D-Caffe. https://github.com/faustomilletari/3D-Caffe. (Accessed: June 2018). 2

[GFJ16] GUEUNET C., FORTIN P., JOMIER J.: Contour forests: Fast multi-threaded augmented contour trees. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)* (2016), pp. 85–92. 1

[ICNK17] IOANNIDOU A., CHATZILARI E., NIKOLOPOULOS S., KOMPATSIARIS I.: Deep Learning Advances in Computer Vision with 3D Data: A Survey. *ACM Comput. Surv. 50*, 2 (2017), 20:1–20:38. 2

[JSD*14] JIA Y., SHELHAMER E., DONAHUE J., KARAYEV S., LONG J., GIRSHICK R., GUADARRAMA S., DARRELL T.: Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia* (New York, NY, USA, 2014), MM '14, ACM, pp. 675–678. 2

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* (2012), NIPS'12, pp. 1097–1105. 1

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. In *Proceedings of the IEEE* (1998), pp. 2278–2324. 1

[LDS*17] LUENGO I., DARROW M., SPINK M., SUN Y., DAI W., HE C., CHIU W., PRIDMORE T., ASHTON A., DUKE E., BASHAM M., FRENCH A.: SuRVoS: Super-Region Volume Segmentation workbench. *Journal of Structural Biology 198*, 1 (4 2017), 43–53. 1

[LTvdV*14] LITJENS G., TOTH R., VAN DE VEN W., HOEKS C., KERKSTRA S., VAN GINNEKEN B., VINCENT G., GUILLARD G., BIRBECK N., ZHANG J., STRAND R., MALMBERG F., OU Y., DAVATZIKOS C., KIRSCHNER M., JUNG F., YUAN J., QIU W., GAO Q., EDWARDS P. J., MAAN B., VAN DER HEIJDEN F., GHOSE S., MITRA J., DOWLING J., BARRATT D., HUISMAN H., MADABHUSHI A.: Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge. *Medical Image Analysis 18*, 2 (2014), 359 – 373. 1

[MMH*17] MIKI Y., MURAMATSU C., HAYASHI T., ZHOU X., HARA T., KATSUMATA A., FUJITA H.: Classification of teeth in cone-beam CT using deep convolutional neural network. *Computers in Biology and Medicine 80* (2017), 24–29. 1

[MNA16] MILLETARI F., NAVAB N., AHMADI S. A.: V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *International Conference on 3D Vision (3DV)* (2016), pp. 565–571. 1, 2, 3, 4

[MRAG08] MOHAMMAD H., REZA A. Z., ALI A. T.-F., GHOLAMREZA S.: Segmentation of Teeth in CT Volumetric Dataset by Panoramic Projection and Variational Level Set. *International Journal of Computer Assisted Radiology and Surgery 3*, 3 (2008), 257–265. 1

[MS15] MATURANA D., SCHERER S.: VoxNet: A 3D Convolutional Neural Network for real-time object recognition. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015), 922–928. 2

[MSH*17] MOLAEI S., SHIRI M., HORAN K., KAHROBAEI D., NALLAMOTHU B., NAJARIAN K.: Deep convolutional neural networks for left ventricle segmentation. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (July 2017), pp. 668–671. 1

[NHH15] NOH H., HONG S., HAN B.: Learning deconvolution network for semantic segmentation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)* (Washington, DC, USA, 2015), ICCV '15, IEEE Computer Society, pp. 1520–1528. 1

[OPT15] OLAF R., PHILIPP F., THOMAS B.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – (MICCAI)* (2015), Nassir N., Joachim H., M. W. W., F F. A., (Eds.), pp. 234–241. 1

[PJSM15] PAUWELS R., JACOBS R. T., SINGER S. R., MUPPARAPU M.: CBCT-based bone quality assessment: are Hounsfield units applicable? *Dentomaxillofacial Radiology 44*, 1 (2015), 20140238. 3

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., BERG A. C., FEI-FEI L.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV) 115*, 3 (2015), 211–252. 1

[SLD17] SHELHAMER E., LONG J., DARRELL T.: Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 4 (April 2017), 640–651. 1

[ZKZ*18] ZHONG Z., KIM Y., ZHOU L., PLICHTA K., ALLEN B., BUATTI J., WU X.: 3d fully convolutional networks for co-segmentation of tumors on pet-ct images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (April 2018), pp. 228–231. 1