This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Let Me DeCode You: Decoder Conditioning with Tabular Data

Tomasz Szczepański[1], Michal K. Grzeszczyk[1], Szymon Płotka[1,2,3], Arleta Adamowicz[4], Piotr Fudalej[4], Przemysław Korzeniowski[1], Tomasz Trzciński[5,6,7], and Arkadiusz Sitek[8]

[1] Sano Centre for Computational Medicine, Cracow, Poland
t.szczepanski@sanoscience.org
[2] Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
[3] Amsterdam University Medical Center, Amsterdam, The Netherlands
[4] Jagiellonian University Medical College, Cracow, Poland
[5] Warsaw University of Technology, Warsaw, Poland
[6] IDEAS NCBR, Warsaw, Poland
[7] Tooploox, Wroclaw, Poland
[8] Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract. Training deep neural networks for 3D segmentation tasks can be challenging, often requiring efficient and effective strategies to improve model performance. In this study, we introduce a novel approach, DeCode, that utilizes label-derived features for model conditioning to support the decoder in the reconstruction process dynamically, aiming to enhance the efficiency of the training process. DeCode focuses on improving 3D segmentation performance through the incorporation of conditioning embedding with learned numerical representation of 3D-label shape features. Specifically, we develop an approach, where conditioning is applied during the training phase to guide the network toward robust segmentation. When labels are not available during inference, our model infers the necessary conditioning embedding directly from the input data, thanks to a feed-forward network learned during the training phase. This approach is tested using synthetic data and cone-beam computed tomography (CBCT) images of teeth. For CBCT, three datasets are used: one publicly available and two in-house. Our results show that DeCode significantly outperforms traditional, unconditioned models in terms of generalization to unseen data, achieving higher accuracy at a reduced computational cost. This work represents the first of its kind to explore conditioning strategies in 3D data segmentation, offering a novel and more efficient method for leveraging annotated data. Our code, pre-trained models are publicly available at https://github.com/SanoScience/DeCode.

Keywords: Conditioning · Tabular data · Non-Imaging · Segmentation

1 Introduction

The annotation process in medical imaging is time-consuming, costly, and requires medical domain knowledge [17]. Furthermore, deep learning-based algo-

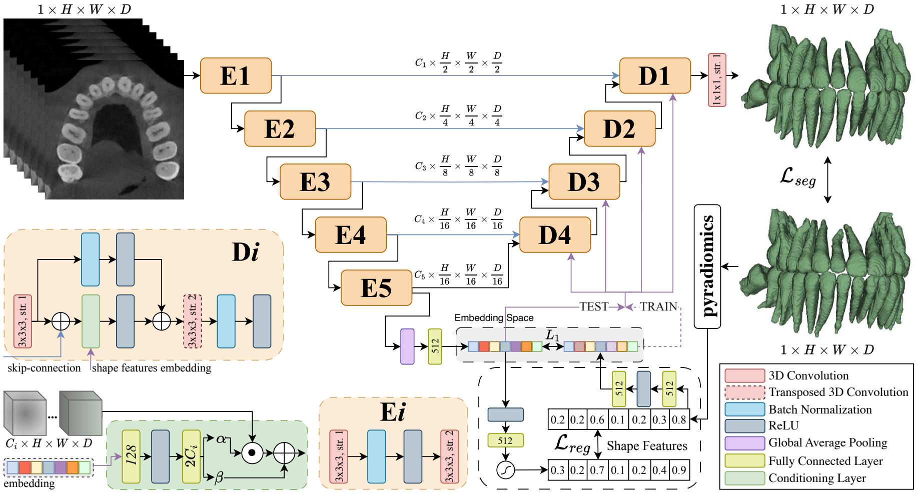

Fig. 1. An overview of the proposed DeCode method for conditioning segmentation decoder with learned shape features embedding. During the inference when test labels are unavailable, we use the learned feature embedding optimized with L_1 loss in Eq. 1. We perform conditioning after the skip connection from the encoder, allowing for a dynamic and selective decoding process. We also leverage a features regression as a helper task that boosts meaningful feature extraction. Skip connections and the flow of the shape features embedding are indicated with blue and purple arrows respectively. Ei and Di correspond to the encoder and decoder stages.

rithms necessitate a large amount of annotated data for acceptable performance and generalization capabilities [8]. However, to enhance deep learning algorithms without relying solely on large-scale imaging data, the community explored the use of tabular features [5,23,13].

In recent years, FiLM [12] has emerged, allowing adaptive influence on neural network intermediate features through feature-wise affine transformations based on conditioning information. Expanding on this concept, integrating tabular information has shown significant advantages for model performance [3,22,6]. An example of the beneficial integration is TabAttention mechanism [3]. This approach incorporates biometric measurements which improve fetal birth weight estimation on ultrasound video scans. Similarly, DAFT [22] is proposed to conditionally shift and scale feature maps based on conditioning. Integration of a patient's clinical information with a 3D MRI image shows improvement in time-to-dementia prediction, underscoring the richness of Electronic Health Record (EHR) data. However, both methods primarily address regression problems where tabular data exhibits measurable correlations with the target task, and imaging features contribute to reducing estimation error.

Conversely, segmentation tasks often lack corresponding tabular data due to challenges such as fully anonymization process of medical data, the absence of a comprehensive data collection strategy, or clear relations between conditioning information and segmentation. The authors of conditioning layer INSIDE [6] propose to integrate non-imaging information into 2D segmentation network to improve performance. They utilize cardiac cycle phase and encoded 2D slice position as conditioning data. However, this prior-knowledge-based information relies on a simple two-state phase of the cardiac cycle, though correlated with the segmentation task, limits its information to a simple binary flag. This attempt sets out a line of research that we choose to pursue.

In this work, we explore conditioning on 3D data in a segmentation task when corresponding tabular data is unavailable. To our best knowledge, we are the first to investigate it. We introduce the DeCode, which performs conditioning based on shape feature embedding. For reproducibility, we calculate the label-based shape features using PyRadiomics [20]. We also demonstrate that the accompanying task of shape features regression benefits the model's segmentation performance and, most importantly, allows us to use feature embedding for conditioning during inference when test labels are unavailable. We evaluate our method on the novel synthetic DeCode 3D dataset, showing that shape features allow for conditioning synthetic tasks, thus demonstrating their usefulness. To demonstrate DeCode's applicability in a clinical setting, we use the 3D dental CBCT dataset [1] to train the model and two external test sets to evaluate the generalization of the model. Accurate tooth delineation in dental CBCT images is essential for clinical diagnosis and treatment while preparing precise 3D labels is very time-consuming [14,24,7,2,21]. Our conditioned architecture improves generalization to unseen data compared to the unconditioned one, which is trained on the same data while requiring no extra labeling work and only marginal additional training time. This work proposes a conditioning strategy for 3D data segmentation, offering a more efficient method for leveraging annotated data.

2 Method

In this section, we provide a detailed description of the network and the DeCode decoder with an emphasis on the application of conditioning information. Then, we describe the process of calculating shape features that are further used for the regression task to generate their embedding for inference-time conditioning.

The go-to standard for accomplishing medical imaging segmentation tasks is U-shaped architectures [10,1,15,9]. Here, we follow this approach and present lightweight architecture with an overview of our method in Fig. 1. Let X be a 3D CBCT scan $X \in \mathbb{R}^{1\times H\times W\times D}$ of height H, width W and depth D, the U-shaped network generates multi-scale features at encoding stages Ei. Deeper encoder stages yield more abstract features up to the bottleneck within the deepest part of the architecture, containing compressed information from the input image. The decoding part Di aims to reconstruct the segmentation map from features extracted by the encoder with additional skip connections at consecutive stages. Starting with high-level features of shape $14\times14\times10$ in the bottleneck through all

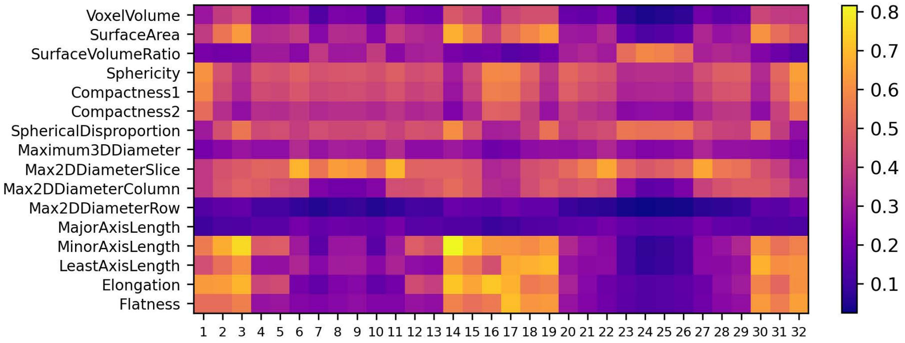

Fig. 2. Normalized mean shape features calculated with PyRadiomics [20] on CBCT Tooth dataset [1]. Each shape feature is calculated for every tooth separately revealing morphological differences between tooth types.

decoder stages, we utilize learned feature embedding to condition the decoding process to improve the quality of the output mask.

Decoder Conditioning. The first step within the decoding step is processing features from the previous stage with the convolutional layer. We add features from the encoder skip connections just before the conditioning layer to avoid leakage of low-level features without first conditioning them on shape feature embedding. The conditioning layer utilizes affine transformations to scale and shift feature maps. The transformation parameters are α_c and β_c, the products of hyper-network, which implement scale and offset, where c is the number of feature map's channels (see Conditioning Layer in Fig. 1). In contrast to FiLM conditioning, we parameterize the scale parameter to $(1 - \alpha_c)$ to facilitate the identity transform especially at the early stage of training, and to allow the scaling parameter to be regularized as a distance from zero. For $\alpha > 0$, the scaling factor inverts a feature map, highlighting features that the ReLU activation would have otherwise suppressed [16]. The conditioning operation takes a normalization role, replacing the batch norm between the convolutional layer and the activation function. In addition to the possibility of conditioning itself, this operation has an additional advantage compared to Batch Norm or Layer Norm: it does not depend on batch statistics [18]. The transformed features are summed via residual connection with the processed input to the decoder. The decoding step finishes with refining and up-sampling feature maps via transposed convolution, Batch Normalization, and a ReLU activation.

Shape Features. We utilize ground-truth masks to extract rich information and shape features. We consider features such as sphericity, volume, and elongation (see Fig. 2). These morphometric descriptors analyze size, form, and shape, and are thus closely linked to the morphology of the segmented objects. Incorporating them aims to decode more morphologically accurate masks. Before training, we extract shape features for every segmented object separately based on the ground truth mask (up to 32 objects, teeth, in a CBCT scan). This process yields a vector of length 512, forming tabular data that is used further to learn conditioning embedding. These features are utilized during training time to condition the decoding process. In real-world scenarios, the ground-truth masks are unavailable during inference. Therefore, we perform a shape-features regression task from the encoder's bottleneck latent space to replace the unavailable shape-features at the test stage with model-learned embedding (see Fig. 1). For reproducibility and easy access, we calculate shape features with PyRadiomics [20].

Loss function. The multi-task loss function, which minimizes both segmentation, embedding distance, and regression tasks, is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathit{\Lambda}_1 \mathcal{L}_{\text{Focal}} + L_1 + \mathit{\Lambda}_2 \mathcal{L}_{\text{RMSE}} + \eta(\|\alpha\|_2^2 + \|\beta\|_2^2), \tag{1}$$

where $\mathit{\Lambda}_1 = 0.5$, and $\mathit{\Lambda}_2 = 0.75$. The coefficients are determined based on a trial-and-error optimization. $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{Focal}}$ correspond to the segmentation task. We optimize an L_1 distance to make encoder features embedding close to the representation of tabular shape features. During inference, we use this learned embedding to condition the decoder. We also add the helper task of shape features regression during training which we optimize based on Root Mean Square Error (RMSE). Finally, we add an L_2 penalty $\eta = 0.00001$ to regularize the conditioning layer parameters α and β, due to the high capacity of the deep network following the conditioning layer, thus reducing the risk of overfitting.

3 Experiments and Results

In this section, we describe implementation details and introduce the synthetic dataset, 3DeCode, where we investigate the possibility of conditioning with shape features on 3D data. Moreover, we apply DeCode to the task of 3D segmentation, utilizing CBCT datasets. We highlight the significance of DeCode key components and evaluate its ability to generalize to unseen CBCT data in comparison to lightweight 3D UNet, which lacks decoder conditioning.

Implementation details. We implement identical models for both synthetic and clinical datasets. We use a UNet network with 4-down and 4-up sampling stages, Batch Normalization, ReLU activations, and a Sigmoid layer for final classification. The conditioning layers are placed inside decoder stages as shown in Fig. 1. We crop an ROI around the teeth based on labels with a size of 240×240×176 from the input CBCT scan. Then, we randomly crop a patch of size 224×224×160 and feed it to the network. We train the model using a batch size of 4, and the AdamW optimizer for 400 epochs. A learning rate is set to 0.001, and the weight decay is set to 0.0001. The intensity of the Hounsfield Unit is clipped to the range [0, 3500] and linearly scaled to [0, 1]. We employ geometric and intensity-related data augmentation such as random rotation, translation, or brightness and contrast adjustments throughout the training process. We implement our model in PyTorch 1.13.1 and MONAI 1.2.0 and train it on NVIDIA A100 80GB GPU with CUDA 11.6. We use PyRadiomics 3.1.0 to calculate binary mask shape features. In case of a missing tooth, we fill its shape features

Table 1. Quantitative results on DeCode 3D dataset in the average Dice Similarity Coefficient (DSC) (%). We explore the possibility of conditioning 3D solid with shape features in a segmentation task. For the tasks *Size* and *Shape*, we conditionally segment solids of small, medium, or large sizes and sphere, cube, or cylinder shapes, respectively. *Mixed* task configurations combine the characteristics of *Shape* and *Size* tasks. More challenging *Varying* combinations additionally address shape and size variability based on a uniform distribution, beyond binary combinations. We report the baseline as an unconditioned UNet.

Task	DSC (%)	
	Baseline	Shape features conditioning
Size	49.18 ± 32.26	98.23 ± 3.92
Shape	53.48 ± 22.09	99.33 ± 0.85
Mixed	17.84 ± 25.68	97.96 ± 5.25
Varying Size	32.97 ± 32.23	97.48 ± 5.69
Varying Mixed	12.43 ± 28.45	94.74 ± 12.94

with a vector of zeros and finally normalize tabular data to a range of [0, 1]. We perform a paired t-test with $p < 0.05$ to identify significant differences.

3DeCode dataset. We present a novel dataset inspired by CLEVR-Seg [6], extending it to 3D and generating segmentation masks based on conditioning scenario tasks. We design tasks that require conditioning based on Shape, Size, or Shapes of different Sizes (referred to as Mixed). To utilize the rich information stored as non-binary shape features, we also enrich the dataset with solids of varying shapes and sizes. Namely, we generate two additional tasks that introduce non-discrete variability in Size or Shape to the solids, based on a uniform distribution, e.g., to generate the varying-size solid class 'small sphere' we vary its radius by ± 20%. While this approach does not reflect the full spectrum of information that shape features can store, it allows us to assess the feasibility of conditioning on 3D data in a segmentation task. The Varying Mixed task consists of shapes varying in size and shape, where, e.g., the base spherical shape can result in an ellipsoid and a cube in a cuboid. The generated solids are binary, as complex image feature extraction is not a concern. Tasks to be solved accurately require the use of the conditioning information by the network. Otherwise, accuracy is reduced to a random guess based solely on the image. The positions of the solids are drawn randomly, whereby they may overlap. We generate 300 labeled conditions for tasks of Size (small, medium, or large) or Shape (sphere, cube, cylinder), and 900 for the Mixed tasks. Data consists of condition-based 3D images with up to 18 objects in volume space of the same size as the patch size used by our model. We generate every possible conditioning combination per image to prevent the model from memorizing image-condition pairs. For evaluation, we split the dataset into training, validation, and testing subsets with a 60:20:20 ratio. 3DeCode samples can be found in the supplementary material.

CBCT dataset. To train our model, we use 98 publicly available 3D dental CBCT scans [1]. We evaluate the segmentation performance on an external test

Table 2. Quantitative results on 3D CBCT datasets: external (Center A and Center B) and validation split. We report DSC and standard deviation. Configuration (1) refers to an unconditioned network serving as a baseline. An upper bound of generalization is provided by configuration (7) conditioned with shape features calculated on test-set masks. The proposed configuration DeCode (8) utilizes during test time learned feature embedding. We conduct a paired t-test to establish statistical significance between the baseline and configuration (7) and (8), denoted by (*) for $p < 0.05$. CL stands for the Conditioning Layer, Reg the Regression task, CR Conditioning Information Representation, and T Test-time conditioning, Rand for Random Features, CSF for test-set Calculated Shape Features (an oracle approach), and LESF for Learned Embedding of Shape Features.

	Configuration				DSC (%)		
	CL	Reg	CR	T	Center A	Center B	VAL
1.	-	✗	-	-	89.67 ± 2.34	94.55 ± 1.16	**95.89 ± 0.84**
2.	-	✓	-	-	91.94 ± 1.56	95.41 ± 1.01	95.67 ± 0.88
3.	FiLM	✗	Rand	✗	89.75 ± 1.94	93.72 ± 1.33	95.86 ± 0.95
4.	FiLM	✗	CSF	✗	92.11 ± 1.79	95.45 ± 1.12	95.59 ± 0.88
5.	INSIDE	✗	CSF	✗	91.16 ± 3.33	95.12 ± 0.99	95.61 ± 0.73
6.	DAFT	✗	CSF	✗	90.61 ± 9.22	95.14 ± 1.05	95.54 ± 0.91
7.	FiLM	✓	CSF	✗	**93.12 ± 1.07***	**95.52 ± 0.92***	95.60 ± 0.93
8.	FiLM	✓	LESF	✓	92.74 ± 1.34*	95.12 ± 0.91*	95.81 ± 0.86

set, comprising 20 CBCT scans obtained from a retrospective study (IRB OKW-623/2022) conducted at two medical centers: Center A (11 scans) and Center B (9 scans). Scans were acquired using the Carestream CS 9600 and i-CAT 17 19, with a slice thickness of 0.15 mm/px and 0.2 mm/px, respectively. The ground truth annotations for the test set were performed by an orthodontist with 5 years of experience, who was verified by another orthodontist with 25 years of clinical practice. We resample all scans to $0.4 \times 0.4 \times 0.4$ mm^3 isotropic resolution.

Results. We present results on 3DeCode dataset in Table 1. According to our dataset-building principle, the baseline UNet cannot segment the image without conditioning. The Mixed task DSC is 17% which is close to a random sample 1 out of 9. The model with decoder conditioning can correctly perform the conditional 3D segmentation task, approaching perfect accuracy for the Shape task with a DSC of 99.23%, and 94.74%, respectively for the most challenging Varying Mixed task. Our model struggles only when overlapping solids, due to random placement solids, are present, which is unrelated to conditioning. The results of the experiment, demonstrate that it is possible to condition in 3D using shape features embedding, which allows us to move on to examine the impact of conditioning on clinical data segmentation.

We compare the unconditioned UNet network's (1) results on the CBCT dataset, which serves as the baseline method, with the proposed DeCode (8) – with numbers in brackets corresponding to configurations presented in Table 2. To find the optimal configuration, we explore the impact of an auxiliary shape

Table 3. Performance comparison on CBCT test sets between baseline unconditioned U-shaped networks and the DeCode method. P denotes the number of parameters, I inference time, and T training time.

Network	P (M)	GFLOPs	I (ms)	T (h)		DSC (%)	
					Center A	Center B	Avg.
UNext [19]	4	46	21	1.5	88.98 ± 3.32	93.37 ± 1.80	90.96 ± 2.64
UNet [15]	25	1880	117	8.5	92.03 ± 1.45	94.41 ± 0.98	93.10 ± 1.24
ResUNet34 [4]	70	2610	101	11	92.28 ± 1.32	95.56 ± 0.99	93.71 ± 1.17
Att-UNet [11]	6	380	127	5	92.66 ± 1.51	95.22 ± 1.06	93.81 ± 1.31
VNet [9]	46	2770	175	13	93.07 ± 0.93	95.42 ± 1.02	94.13 ± 0.97
DeCode	4	204	41	3	92.74 ± 1.34	95.12 ± 0.91	93.81 ± 1.15

feature regression task (2), different conditioning layers (CL) (4-6), and conditioning information representation (CR) (3-8). Firstly, we add a shape feature regression task (2) to the baseline method that improves generalization on both external sets, proving the usefulness of the shape features. Secondly, we evaluate conditioning layer types with calculated shape features (CSF), which, for this experiments, we also use during the test (an oracle approach). We get the best results with the FiLM layer, so we use it for further experiments. We examine the edge case of conditioning on random tabular data generated from a standard normal distribution and observe a significant performance decline. The result for Center A is better than the unconditioned model, suggesting that the conditioning layer increases the model's capacity, posing a threat of overfitting. However, it may also suggest that residual connections in the decoder make the model robust to the possible negative impact of conditioning. A final experiment (7) based on the CSF leverages the FiLM layer and regression task. It sets an upper bound for generalization improvement. To adapt the method to test time, unlike configuration (7), we use learned embedding of shape features (LESF), which is our proposed configuration (8). Although the proposed DeCode does not improve the result on the validation set, it statistically significantly improves the generalization to new unknown data. Finally, we compare out method with unconditioned U-shaped networks (see Table 3). We choose architectures with a broad range of parameter numbers, provided they allow training with a large 3D patch under GPU memory constraints. Our solution achieves the second-best generalization, giving way only to the VNet method, which is, however, 10× more computationally intensive and requires 4× longer training.

4 Conclusions

This paper investigates the possibility of conditioning the decoder in the 3D segmentation task on the tabular data. Compared to unconditioned training, DeCode performs better on unseen data, requiring no extra labeling work and marginal additional training time. We evaluated our method on two external

CBCT datasets, proving its enhanced generalizability. Obtained results encourage further research in this field, allowing more efficient use of annotated data.

There are limitations to our method. Firstly, we train our method on a relatively small dataset where selecting hyperparameters is complex, and their small changes may lead to a loss of stability in embedding learning, including their collapse. We expect better stability and further segmentation improvement with the increased dataset. Secondly, the radiomics features provide information limited to shape without considering objects' positions and relations between them. In the future, we plan to conduct the conditioning on features extracted automatically from labels, enabling the end-to-end training of representations for improved clinical image segmentation. Finally, we leverage the consistent shapes of physiological structures like teeth, but DeCode may face constraints with unpredictably shaped structures like tumors.

Acknowledgments. This work is supported by the EU's Horizon 2020 programme (grant no. 857533, Sano) and the Foundation for Polish Science's International Research Agendas programme, co-financed by the EU under the European Regional Development Fund and the Polish Ministry of Science and Higher Education (contract no. MEiN/2023/DIR/3796). This research was funded in whole or in part by National Science Centre, Poland 2023/49/N/ST6/01841. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Cui, Z., Fang, Y., Mei, L., Zhang, B., Yu, B., Liu, J., Jiang, C., Sun, Y., Ma, L., Huang, J., et al.: A fully automatic ai system for tooth and alveolar bone segmentation from cone-beam ct images. Nature Communications **13**(1), 2096 (2022)
2. Cui, Z., Zhang, B., Lian, C., Li, C., Yang, L., Wang, W., Zhu, M., Shen, D.: Hierarchical morphology-guided tooth instance segmentation from cbct images. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. pp. 150–162. Springer (2021)
3. Grzeszczyk, M.K., Płotka, S., Rebizant, B., Kosińska-Kaczyńska, K., Lipa, M., Brawura-Biskupski-Samaha, R., Korzeniowski, P., Trzciński, T., Sitek, A.: Tabattention: Learning attention conditionally on tabular data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 347–357. Springer (2023)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
5. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine **3**(1), 136 (2020)

6. Jacenków, G., O'Neil, A.Q., Mohr, B., Tsaftaris, S.A.: Inside: steering spatial attention with non-imaging information in cnns. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. pp. 385–395. Springer (2020)
7. Li, P., Liu, Y., Cui, Z., Yang, F., Zhao, Y., Lian, C., Gao, C.: Semantic graph attention with explicit anatomical association modeling for tooth segmentation from cbct images. IEEE Transactions on Medical Imaging **41**(11), 3116–3127 (2022)
8. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
9. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
10. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3523–3542 (2021)
11. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
12. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
13. Płotka, S., Grzeszczyk, M.K., Brawura-Biskupski-Samaha, R., Gutaj, P., Lipa, M., Trzciński, T., Išgum, I., Sánchez, C.I., Sitek, A.: Babynet++: Fetal birth weight prediction using biometry multimodal data acquired less than 24 hours before delivery. Computers in Biology and Medicine **167**, 107602 (2023)
14. Polizzi, A., Quinzi, V., Ronsivalle, V., Venezia, P., Santonocito, S., Lo Giudice, A., Leonardi, R., Isola, G.: Tooth automatic segmentation from cbct images: a systematic review. Clinical Oral Investigations **27**(7), 3363–3378 (2023)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
16. Rupprecht, C., Laina, I., Navab, N., Hager, G.D., Tombari, F.: Guide me: Interacting with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8551–8561 (2018)
17. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X.: Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis **63**, 101693 (2020)
18. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 5314–5321 (2022)
19. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 23–33. Springer (2022)
20. Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), e104–e107 (2017)

21. Wang, Y., Xia, W., Yan, Z., Zhao, L., Bian, X., Liu, C., Qi, Z., Zhang, S., Tang, Z.: Root canal treatment planning by automatic tooth and root canal segmentation in dental cbct with deep multi-task feature learning. Medical Image Analysis **85**, 102750 (2023)
22. Wolf, T.N., Pölsterl, S., Wachinger, C., Initiative, A.D.N., et al.: Daft: a universal module to interweave tabular data and 3d images in cnns. NeuroImage **260**, 119505 (2022)
23. Xia, Y., Chen, X., Ravikumar, N., Kelly, C., Attar, R., Aung, N., Neubauer, S., Petersen, S.E., Frangi, A.F.: Automatic 3d+ t four-chamber cmr quantification of the uk biobank: integrating imaging and non-imaging data priors at scale. Medical Image Analysis **80**, 102498 (2022)
24. Zheng, Q., Gao, Y., Zhou, M., Li, H., Lin, J., Zhang, W., Chen, X.: Semi or fully automatic tooth segmentation in cbct images: a review. PeerJ Computer Science **10**, e1994 (2024)