frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 24 January 2024
DOI 10.3389/fphy.2024.1257512

Check for updates

OPEN ACCESS

EDITED BY
Michela Fratini,
National Research Council (CNR), Italy

REVIEWED BY
Bernhard Hesse,
European Synchrotron Radiation Facility, France
Lorenzo Massimi,
University College London, United Kingdom

*CORRESPONDENCE
André Lopes Marinho,
andre.lopes@hereon.de
Berit Zeller-Plumhoff,
berit.zeller-plumhoff@hereon.de

RECEIVED 12 July 2023
ACCEPTED 02 January 2024
PUBLISHED 24 January 2024

CITATION
Lopes Marinho A, Kazimi B, Ćwieka H, Marek R, Beckmann F, Willumeit-Römer R, Moosmann J and Zeller-Plumhoff B (2024), A comparison of deep learning segmentation models for synchrotron radiation based tomograms of biodegradable bone implants.
Front. Phys. 12:1257512.
doi: 10.3389/fphy.2024.1257512

A comparison of deep learning segmentation models for synchrotron radiation based tomograms of biodegradable bone implants

André Lopes Marinho[1]*, Bashir Kazimi[1], Hanna Ćwieka[1], Romy Marek[2], Felix Beckmann[3], Regine Willumeit-Römer[1], Julian Moosmann[1,3] and Berit Zeller-Plumhoff[1]*

[1]Helmholtz-Zentrum Hereon GmbH, Institute of Metallic Biomaterials, Geesthacht, Germany , [2]Department of Orthopaedics and Traumatology, Medical University of Graz, Graz, Austria , [3]Helmholtz-Zentrum Hereon GmbH, Institute of Materials Physics, Geesthacht, Germany

Introduction: Synchrotron radiation micro-computed tomography (SRμCT) has been used as a non-invasive technique to examine the microstructure and tissue integration of biodegradable bone implants. To be able to characterize parameters regarding the disintegration and osseointegration of such materials quantitatively, the three-dimensional (3D) image data provided by SRμCT needs to be processed by means of semantic segmentation. However, accurate image segmentation is challenging using traditional automated techniques. This study investigates the effectiveness of deep learning approaches for semantic segmentation of SRμCT volumes of Mg-based implants in sheep bone ex vivo.

Methodology: For this purpose different convolutional neural networks (CNNs), including U-Net, HR-Net, U^2-Net, from the TomoSeg framework, the Scaled U-Net framework, and 2D/3D U-Net from the nnU-Net framework were trained and validated. The image data used in this work was part of a previous study where biodegradable screws were surgically implanted in sheep tibiae and imaged using SRμCT after different healing periods. The comparative analysis of CNN models considers their performance in semantic segmentation and subsequent calculation of degradation and osseointegration parameters. The models' performance is evaluated using the intersection over union (IoU) metric, and their generalization ability is tested on unseen datasets.

Results and discussion: This work shows that the 2D nnU-Net achieves better generalization performance, with the degradation layer being the most challenging label to segment for all models.

KEYWORDS

synchrotron radiation micro-computed tomography, biodegradable materials, magnesium-based alloys, semantic segmentation, convolutional neural networks

1 Introduction

Synchrotron radiation micro-computed tomography (SRµCT) is a powerful technique to characterize a plethora of different materials non-invasively in 3D [1–3] Magnesium (Mg)-based alloys are one such material class that is increasingly researched using SRµCT [4]. Mg-based alloys are researched in particular as novel materials as bone implants and as stents, because of their biocompatibility and biodegradability [5–7]. The 3D and high-resolution nature of SRµCT in this context enables studying both the material microstructure, as well as their integration into the surrounding tissue and morphology. Quantities of interest include degradation rates (DR), bone-to-implant contact (BIC), and relative bone volume (BV/TV) [8–12]. In order to extract such quantitative information from the 3D images, a prior image segmentation, i.e., a pixel-/voxel-wise classification, is required [13, 14]. However, in some cases, this step represents a major bottleneck, as mapping the structures into labels through the image greyscales is difficult with standard automated techniques [15]. In such cases, machine and deep learning algorithms, specifically convolutional neural networks (CNN) can be employed [10, 16–22].

Over time, several CNN architectures have been developed and refined, leading to significant advancements in deep learning [23–25]. The 'U-Net', introduced by Ronneberger et al. [26], was specifically designed for biomedical image analysis and adopts the Fully Convolutional Network concept proposed by Long et al. [27]. The U-Net architecture has demonstrated promising results across a wide range of semantic segmentation tasks [28–31]. This success has sparked further developments, including the framework 'Scaled U-Net' by Baltruschat and Ćwieka et al. [32], which includes a nine-axis prediction fusing feature for the specific use of segmenting biodegradable Mg-based bone implants. The 'nnU-Net' [33] is a self-configuring U-Net-based framework with automized preprocessing, network architecture, training, and post-processing for any two-dimensional (2D) or 3D segmentation task. The nnU-Net has achieved top rankings in challenges such as the Medical Segmentation Decathlon [34] and the International Symposium on Biomedical Imaging (ISBI) Challenge. The 'U^2-Net', developed by Qin et al. [35], has a nested U-Net architecture with partial encoders for the detection of salient objects in image data. It is designed to address the limitations of existing deep learning architectures for semantic segmentation and was also used as a method for the segmentation of biomedical data [36]. This architecture has been evaluated on various benchmark datasets and compared with other state-of-the-art methods, including the original U-Net architecture. Previous studies have shown that the U^2-Net outperforms the U-Net in terms of segmentation accuracy, particularly for images with complex structures and fine details [35, 37]. Another common architecture is the 'HR-Net' which was initially designed to address the problem of human pose estimation [38]. In this sense, the model proposes the processing of high-resolution images by blocks that connect high-to-low resolutions in parallel. The HR-Net has achieved good results for semantic segmentation in various applications and domains, including state-of-the-art performance on the LIP [39] and Cityscape [40] datasets.

This study focuses on exploring and comparing the application of deep learning techniques for the semantic segmentation of *ex vivo* SRµCT volumes of Mg-based implants in sheep bone. Specifically, screws made of ZX00, a Mg alloy containing < 0.5 wt% Zn and < 0.5 wt% Ca, were implanted in the diaphysis, epiphysis, and metaphysis of sheep tibiae [11]. SRµCT imaging was conducted following sacrifice of the animals after healing periods between 4 and 24 weeks. To extract the required quantitative information, the data sets should be classified into residual (metallic) material, implant degradation layer, mineralized bone tissue, and background. The U-Net, HR-Net, U^2-Net, Scaled U-Net, and the 2D and 3D nnU-Net models are trained and validated against ground truth data. The training and validation process of the U-Net, HR-Net, U^2-Net are carried out using the TomoSeg framework [41]. The intersection over union (IoU) metric, a commonly used benchmark in such cases [42], is employed to evaluate the performance of the models. Subsequently, the trained models are tested on unseen datasets. Furthermore, the predicted results are used to calculate degradation and osseointegration parameters, allowing for the determination of the relative error between the ground truth and predicted segmentations.

TABLE 1 Overview of samples on which SRµCT image acquisition was performed at P05 and P07.

Implantation site	Healing time/ weeks	# of samples	Beamline
Epiphysis	4	1	P07
	6	2	
	12	2	
Metaphysis	4	1	
	6	1	
	12	1	
Diaphysis	6	3	P05
	12	3	
	24	3	

2 Materials and methods

2.1 Animal experiments

The experimental methods are described in more detail in Marek and Ćwieka et al. [11].

2.1.1 Material production

The ZX00 alloy was made under a shielding gas atmosphere using ultra-high-purity Mg, Zn, and, with a nominal composition of < 0.5 wt% Zn and < 0.5*wt.*% Ca at 750°C, as described in Holweg et al. [43]. Rods with a diameter of 6 mm were indirectly extruded at 345°C. The material was produced by ETH Zürich and Cavis AG (Dübendorf, Switzerland). ZX00 screws measuring 16 mm in length and 3.5 mm in diameter were manufactured by Wittner (Ernst Wittner GmbH, Vienna, Austria). The process was performed without lubrication and with polycrystalline diamond tools to avoid cross-contamination and

TABLE 2 Overview of the SRμCT scanning parameters for both beamlines.

Parameters	Beamline	
	P05	P07
Energy/keV	50	60
Exposure time/ms	120	200
Sample-detector distance/mm	40	400
Camera/pixels x pixels	CCD (7,920 × 6,004)	CCD (6,144 × 6,144)
Number of projections	10,001	10,400
Dimension of acquired projections/pixels	7,920 × 3,801	6,144 × 2,701
Effective pixel size/µm	0.92	1.27
Pixel size after 3x binning/µm	2.76	3.79
Field of view/mm	H: 7.3	H: 7.8
	V: 3.5	V: 3.4
Dimensions after reconstruction/pixels	2000 × 2000; 2,500 × 2,500; 2,640 × 2,640	3,000 × 3,000

TABLE 3 Overview of the datasets used in this work. The SRμCT scans were divided into different sets, which served as input to the CNN architectures for training/validation, testing, and prediction.

Dataset type	Implantation site	Healing time	# of samples
Training/ validation	Epiphysis	4	1
		6	1
		12	2
	Metaphysis	4	1
		6	1
Testing	Epiphysis	6	1
	Metaphysis	12	1
Prediction	Diaphysis	6	3
		12	3
		24	3

corrosive attacks. Subsequently, the screws were cleaned in an ultrasonic bath with acetone and dried at room temperature in a clean-room atmosphere. The screws were sterilized using gamma radiation at a dose of 29.2 kGy [43].

2.1.2 Animal experiments

The animal trials (Permit Numbers: BMWFW-66.010/0073-WF/V/3b/2015 and BMBWF-66.010/0107-V/3b/2019) that served as the basis for this work were conducted with the approval of the Austrian Federal Ministry for Science and Research. The work adhered to the guidelines outlined by the European Convention for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes. In this trial, ZX00 screws were implanted into the tibiae of sheep. Surgical procedures involved creating incisions of approximately 2–3 cm in the animal's skin at the diaphysis, distal medial epiphysis, and metaphysis regions. After healing time intervals of 4, 6, 12, and 24 weeks, the animals were euthanized, and their respective tibiae containing the screws were dissected from the proximal, shaft, and distal parts [11].

2.1.3 *Ex vivo* SRμCT data acquisition

The SRμCT images were acquired at the μCT endstations of the P05 imaging beamline (IBL) [44] and the P07 high energy material science (HEMS) beamline [45]. The beamlines are operated by Hereon at the PETRA III storage ring at the Deutsches Elektronen-Synchrotron (DESY) in Hamburg, Germany. An overview of the imaged samples at the aforementioned beamlines is shown in Table 1. Moreover, Table 2 summarizes key scanning parameters from the experiments at both beamlines. Imaging was performed by rotating off-center ~ 360°. At both endstations, a scintillator made of cadmium tungstate (CdWO4) was used to convert the incoming X-rays to optical light, which was further magnified by an objective lens (×5 for the samples imaged at P05 and 10x for the samples imaged at P07). The visible spectrum was then detected by cameras with a CMOS (complementary metal–oxide–semiconductor) sensor. The tomographic reconstruction was performed in a MATLAB (The MathWorks Inc., USA) framework [46, 47]. For tomographic reconstruction, the filtered back-projection (FBP) algorithm was employed using the ASTRA toolbox for back-projection [48, 49].

2.1.4 Preparation of the SRμCT data

The SRμCT volumes were divided into three groups: training/validation, testing, and prediction. Information about each of these groups can be seen in Table 3. The samples from the epiphysis and metaphysis explants were used for training and validation, and testing, while the diaphysis explants were used for prediction. In this sense, the training and validation dataset contained a total of six samples from different healing times so that the CNNs could be trained considering the different degradation stages of the Mg implants. The testing dataset was composed of two samples that were not used for training and validation to prevent a biased

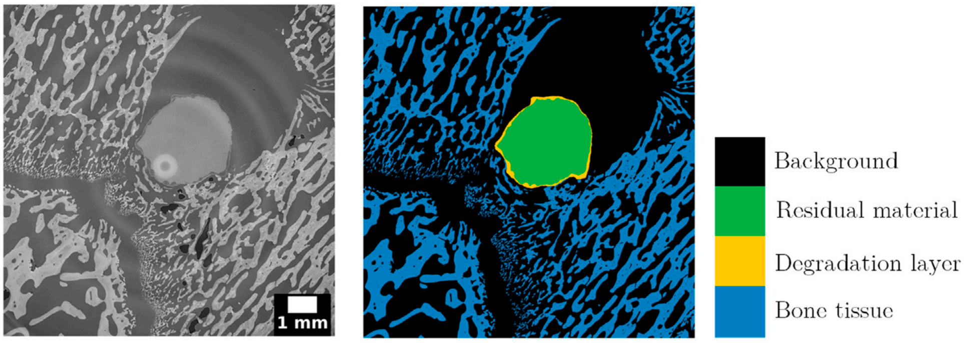

FIGURE 1
Selected cross-section of SRμCT volume and corresponding segmentation. The datasets used for training/validation and testing were segmented into four labels: residual material, degradation layer, bone tissue, and background. The bright circle in the middle of the SRμCT cross-section is an artifact from the image stitching process for 360° reconstruction. Furthermore, ring artifacts are clearly visible in the top left background area.

prediction in the testing phase. Prior to training/testing/validation, all the datasets were denoised with an iterative non-local means filter [50] and the grayscale values were linearly normalized to a [0, 1] range. An example of the normalized values for each label can be seen in Supplementary Appendix Figure SA1, for data from each beamline. Moreover, image quality metrics such as signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) were calculated to compare the training/validation and test datasets with the datasets used for prediction, since the images originated from different sources, i.e., with different experiment setups. These results can be seen in Supplementary Appendix Tables SA1, SA2.

2.1.5 Segmentation of the ground truth data

The ground truth data of the datasets for training/testing and validation was obtained using the software Avizo 2021.1 (FEI SAS, Thermo Scientific, France). The ZX00 volumes were segmented into four labels, namely, residual material (non-corroded alloy), degradation layer (corrosion products around residual material), bone tissue (mineralized tissue), and background (all remaining features). An illustrative example of the segmentation can be visualized in Figure 1. This process was done semi-automatically using the watershed algorithm to obtain the residual material and degradation layer labels, which were then locked to obtain the bone tissue label through an automatic threshold. All the segmentations went through manual corrections to obtain reliable information for the quantifications which will be further calculated. It is important to point out that these segmentations can only approximate the ground truth. Therefore, we expect scores below 100% (as long as overfitting is avoided), in particular for the degradation layer which has the greatest uncertainties.

2.2 Training/validation of the CNN models

Six different CNN architectures from three frameworks were trained to compare the performance of these models for the semantic segmentation of the SRμCT volumes. All models were trained and validated using a 4-fold cross-validation [51]. The training/validation of the 2D implementations of the U-Net, the HR-Net, and the U^2-Net were performed using the framework TomoSeg [41, 52]. Initially, different image input sizes (512 × 512, 768 × 768, and 1,024 × 1,024) were tested within this framework to analyze the influence on the segmentation performance, resulting in the evaluation of nine models (three models for each architecture). Based on the training/validation mIoU of the TomoSeg, an image size of 1024 × 1024 was selected for further comparison with the other classification models. Further modifications to the base structure of the implementations (e.g., of the hyperparameters) were not performed. For this framework, each training/validation lasted approximately 26 h, using two NVIDIA V100 GPUs.

The U-Net implementation of Scaling the U-net framework [32] was trained with an image input size of 512 × 512 . This image size was selected since Baltruschat et al. [32] had shown that these dimensions achieved better segmentation performance in their study. Furthermore, this implementation uses nine-axis fusing which limits the training/validation dimension of the 3D volume to the lowest dimension of the datasets used in this study, i.e., 625 pixels. Further modifications to the base structure of the implementation (e.g., of the hyperparameters) were not performed. The training was completed after 22 h, using two NVIDIA V100 GPUs.

Finally, the nnU-Net [33] framework was used to train/validate both the 2D and 3D U-Net implementations. For training of the considered models, the parameters input size, loss function, and optimizer are automatically selected by the framework, given the datasets used and the labels to be classified. The training process for the 2D U-Net took approximately 21 h, using two NVIDIA V100 GPUs. By contrast, the training of the 3D U-Net took approximately 45 h, using 10 NVIDIA RTX8000 GPUs.

2.3 Testing of the CNN models

2.3.1 mIoU evaluation

Two datasets already segmented were used as references to test the model's predicted segmentation and evaluate them through their mIoU scores. In this context, after predicting the segmentation from the previously trained CNNs, the intersection over union (IoU) of all labels was calculated with Eq. 1 [53]:

$$IoU = \frac{(A \cap B)}{A \cup B} = \frac{TP}{TP + FP + FN} \tag{1}$$

where A is the ground truth segmentation and B is the predicted segmentation. This calculation can also be described in terms of TP - true positives, FP - false positives, and FN - false negatives. From the results, the mean intersection over union (mIoU), which is the average between the IoU obtained from each label (residual material, degradation layer, bone tissue, background) was calculated as (Eq. 2):

$$mIoU = \frac{IoU\,(\text{res. mat.}) + IoU\,(\text{deg. layer}) + IoU\,(\text{bone tissue}) + IoU\,(\text{background})}{4} \tag{2}$$

2.3.2 Quantification of degradation and osseointegration parameters

The segmentations predicted by the CNNs were used to calculate the degradation and osseointegration parameters DR [mm/a], BIC [%], and BV/TV [%]. The same was performed on the ground truth segmentations and the relative errors between the ground truth and predicted values were calculated with Eq. 3.

$$\text{relative error}\,[\%] = \frac{\text{parameter}\,(\text{prediction}) - \text{parameter}\,(\text{ground truth})}{\text{parameter}\,(\text{ground truth})} \tag{3}$$

The DR was calculated from Eq. 4 [54], which involves the volume loss (VL) of the implant over time and characterizes the implant degradation:

$$DR = \frac{V_i - V_r}{A_i \cdot t} = \frac{VL}{A_i \cdot t} \tag{4}$$

V_i is the initial screw volume (reference volume) and V_r is the residual screw volume. A segmented volume of a reference screw is used to calculate V_i. Since both the reference and degraded screw sample must represent the same volume section, a reference implant is registered and resampled on the predicted segmentations volume. A_i is the initial surface reference area and it is given by the total number of surface voxels of the reference screw multiplied by the voxel face area. t is the degradation time.

To evaluate the osseointegration of the implant and consequently its stability, the BIC was calculated according to Eq. 5 [9–11].

$$BIC = \frac{\#\,\text{boundary voxel faces of implant in contact with bone}}{\#\,\text{total surface voxel faces of implant}} \tag{5}$$

The number of boundary voxel faces of the implant in contact with the bone is calculated by counting the number of voxels from the residual material + degradation layer labels in contact with the bone tissue label. Similarly, the total number of surface voxel faces of the implant is given by the number of voxels from the residual material + degradation layer labels in contact with the background label and bone tissue label. The BV/TV was calculated according to Eq. 6 [9–11], making it possible to analyze the bone formation around the implant.

$$BV/TV = \frac{\#\,\text{bone voxels in ROI}}{\#\,\text{bone + background voxels in ROI}} \tag{6}$$

For this purpose, one region of interest (ROI) around the implant of each sample was selected by enlarging the non-degraded, registered reference screw. The ROI corresponded to an enlargement of 1 mm, which is approximately twice the size of the screw threads. The size was selected to account for the effect the threads might have on bone tissue formation.

All the calculations needed for the characterization of the ZX00 implants were performed on the segmented data and were computed through a Python script, in which libraries including NumPy [55], SciPy [56], and sci-image [53] were used to perform image manipulation and processing.

2.4 Predictions of unlabeled data and further comparisons

After selecting the best model among those evaluated in this work, nine unlabeled SRμCT volumes from the prediction dataset served as input for obtaining their segmentations through deep learning. Thus, the generalization performance of the model was assessed.

3 Results and discussion

3.1 Comparison of the CNN models performance

3.1.1 Training and validation—influence of image input size on mIoU

Figure 2A shows the obtained results for the training and validation mIoU for the TomoSeg framework, considering different image input sizes. Overall for this framework, the mIoU improved with increasing of the image input size. The HR-Net showed the best performance, obtaining a higher mIoU (90.95%) for an image input size of 1,024 × 1,024. For the same input size, the U-Net had a similar performance (90.88%). In general, the U^2-Net performed inferiorly to the other models, except for the train image size of 512 × 512. The preparation of the input data for training and validation involves the pre-processing of the volumes by patching the input image to the required size of the model's first convolutional layer. Since this cropping is done randomly, a larger input size means less information loss in training and validation and leads to an increase in the probability that the input image will be patched with information concerning all the classes to be identified, increasing the mIoU. Due to implementations of the used frameworks, a non-random, targeted patch selection strategy to more directly address the challenge of under-represented classes, was not possible, directing the focus toward prioritizing the optimization of patch size. Figure 2B shows the mIoU comparison for the best-performing TomoSeg implementations (those with an image input size of 1,024 × 1,024) and the other trained frameworks (Scaled U-Net, and 2D and 3D nnU-Net), where the image input size was automatically chosen by the application.

The results displayed in Figure 2B show that the best training and validation performance in terms of mIoU was obtained by the 2D nnU-net (mIoU = 94.95%), followed by the 3D nnU-Net (mIoU = 93.10%). The Scaled U-Net obtained a mIoU performance equal to HR-Net, while the U^2-Net continued to

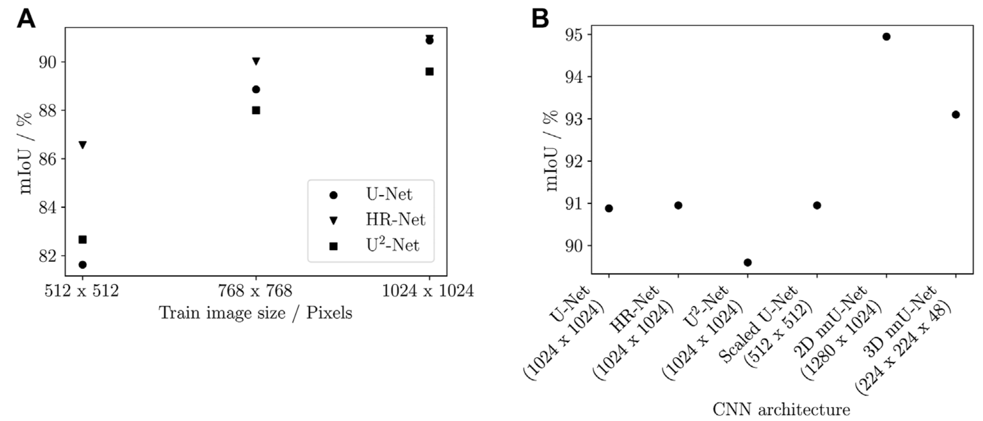

FIGURE 2
Effect of the image input size on the training/validation mIoU of the different models. **(A)** Effect of image input size on mIoU in the TomoSeg framework. The U-Net, HR-Net, and U^2-Net model implementations were trained with different image input sizes to verify the influence of this parameter on the segmentation performance. **(B)** Considering the best-performed models within the TomoSeg framework (image input size of 1,024 × 1,024), further comparisons were made with the Scaled U-Net and nnU-Net frameworks. For the Scaled U-Net, the image size was selected according to the results by Baltruschat et al. [32]. The input sizes for the nnU-Net models were automatically chosen by the framework.

have an inferior segmentation performance among the compared models. Unlike the TomoSeg framework, the training and validation of the Scaled U-Net and the nnU-net considered one image input size. Comparing the effect of image input size on mIoU, Baltruschat et al. [32] obtained better performance for a comparable segmentation task with an image input size of 512 × 512, which was also chosen for this work. Despite using a smaller input size and also being a 2D model, the Scaled U-Net incorporates 3D information through multi-axes fusing, which includes more information for model training, using techniques such as rotating the analyzed volumes around each of its axes [32]. This may explain, therefore, the better performance of this model compared to HR-Net and U-Net, with image input sizes of 1,024 × 1,024. Concerning the nnU-Net, the input image sizes for the two models analyzed were optimized automatically by the framework. In this case, the implementation prioritizes large patch sizes so that the contextual information during training and validation is increased. This is done by considering factors such as the original voxel size of the data and the ratio of classes that are labeled [33]. For the 2D model, it can be seen that the used value of 1,280 × 1,024 is closer to the image input size values from the best-performing models obtained with the TomoSeg framework. The 3D model, however, uses an input size of 228 × 228 × 48, especially since the automation decisions of nnU-Net consider factors such as the limited memory budget of the GPU in use, which is the case when training 3D models. In fact, while all 2D models in this work were trained with 2 GPUs and the training time was between 21 and 26 h, the 3D model in the nnU-Net took 10 GPUs and the training time was 45 h. In this context, it can be seen that although the 3D model of nnU-Net performs well with respect to semantic segmentation, robust hardware is required for the architecture to be trained. For users who do not have the availability of GPUs with full computing power, 2D models are more feasible in terms of hardware and time.

The mIoU results of the trained models are influenced by factors such as the architecture of the models, the number of downsampling and resolution stages, the activation functions, and the normalization techniques used. These factors impact the feature extraction, weight backpropagation, and overall convergence of the models, thus affecting the resulting mIoU scores. For the TomoSeg framework, the implementations of the U-Net, the HR-Net, and the U^2-Net followed the base architecture of each model proposed in the literature and, therefore, will not be discussed in much detail. The framework used convolution operations followed by batch normalization and ReLU activation. U-Net had four downsampling operations for feature extraction. HR-Net had four stages with a new resolution block added from the second stage onwards. U^2-Net had five encoder and four decoder stages with RSU blocks. Scaled U-Net had an extra resolution level compared to TomoSeg's U-Net. Mish function was used for activation after batch normalization. The additional resolution layer, along with the nine-axis fusion implementation, contributed to achieving similar mIoU as higher input sizes trained with the TomoSeg framework for the 512 × 512 input size in the 'Scaled U-Net' implementation. The nnU-Net automatically configures itself based on the input dataset. While it shares the general structure of the U-Net model, some modifications were made, including the use of 8 different resolution levels for both trained models in the framework. Moreover, to enable larger patch sizes, Isensee et al. [33] proposes the use of instance normalization [57], while the other models studied in this work used batch normalization. According to the literature, batch normalization has inferior performance when large patch sizes are considered [57, 58]. Moreover, the activation function in both nnU-Net models is based on the leaky ReLU, which considers a negative slope for values smaller than zero. This allows for feature maps to be weighted also for negative values, influencing the weight backpropagation. In literature, CNNs that used leaky ReLU showed faster convergence than those that used only ReLU [59, 60].

3.1.2 Testing data—mIoU of predicted test samples

The test of the trained models was performed by using the data volume of two samples that were not used in the training process. Only

TABLE 4 mIoU values for the tested CNN models. Tests were performed by predicting the segmentation of two unseen sample data (from 6 to 12 weeks of implantation times). Bold text shows the highest values for each column. *Models from TomoSeg framework with image input size of 1,024 × 1,024.

Model	Intersection over union/%									
	Background		Residual material		Degradation layer		Bone tissue		mIoU/%	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
U-Net*	95.40	97.79	98.84	99.40	84.66	61.17	98.11	96.26	94.26	88.66
HR-Net*	92.54	97.35	98.77	99.38	84.28	65.59	96.87	95.29	93.12	89.40
U^2-Net*	**96.21**	97.87	97.41	99.23	79.48	60.13	**98.57**	96.44	92.92	88.42
Scaled U-Net	90.89	96.36	97.89	97.22	79.57	60.53	94.03	93.33	90.09	87.36
2D nnU-Net	94.47	**98.28**	**98.94**	**99.48**	**86.59**	**71.18**	97.73	**97.09**	**94.43**	**91.49**
3D nnU-Net	93.32	98.27	**98.94**	**99.48**	86.40	70.99	97.24	97.08	93.98	91.46

the models compared in Figure 2B were considered for testing. Table 4 shows for each tested model, the IoU for the segmented labels of each sample and their respective mIoU. In general, the labels background, residual material, and bone tissue showed IoUs above 90% for both samples, in all considered models. However, the degradation layer proved to be the most difficult to segment, obtaining the lowest IoUs. For sample 1, the degradation layer achieved IoUs between 79.48% (for the U^2-Net) and 86.59% (for the 2D nnU-Net). For sample 2, the U^2-Net also had the lowest accuracy for the degradation layer between the models, reaching an IoU of 60.13%., while the 2D nnU-Net had the best IoU between the models for the same label (IoU = 71.18%). Baltruschat et al. [32] reported similar findings for the degradation layer segmentation in a comparable segmentation task. The study reported lower IoU values for the label degradation layer (79.31%–80.16%) than for the classes residual material and bone tissue, which achieved substantially higher IoU values (93.65%–93.10% and 96.72%–96.83%, respectively). Moreover, in a previous study using similar *ex vivo* data, Bockelmann et al. [17] showed that the label ”corroded screw” (degradation layer in this work) was the most challenging to be predicted, achieving a maximum Dice score of 54.1%. The Dice score is a metric similar to the IoU and also measures the performance of the semantic segmentation of ground truth data in comparison with predicted data. In general, the difficulty of the models in segmenting the degradation layer is related to the grayscale values present in the tomographic reconstructions for this region. In fact, the degradation layer has a more heterogeneous appearance due to the different corrosion products and discontinuities such as cracks in some parts of this region. This makes it difficult for the CNN models to generalize the predictions of this label since each sample presents peculiarities in the morphology of this layer. In the broader context, the degradation layer contains a smaller voxel count (surface-to-volume ratio) compared to the residual material, bone tissue, and background labels. This discrepancy in voxel distribution introduces a fairness challenge in the segmentation of the degradation layer. Given the less prominent nature of the corrosion phase, more voxels corresponding to the corrosion phase can significantly impact the results, unlike the other phases where such variations might be less consequential [61]. Since the models trained in this work make a random initial patch of the input data, this problem is further escalated due to the probability of a small number of voxels corresponding to the degradation layer being present in the training/validation phases. It is important to note that the choice to use semi-automatic segmented CT volumes as ground truth was pragmatic, considering their practical availability despite potential limitations. While a phantom volume could offer more controlled ground truth, our focus was on assessing current frameworks within the context of available real-world data.

3.1.3 Qualitative and quantitative analysis of predicted segmentation of test samples

To better interpret the differences in IoU, Figure 3 displays example slices for the segmentations of each tested model for sample 1. As can be seen, the degradation layer is proportionally the label with the least information in the whole segmentation. The differences between the degradation layer label from the ground truth segmentations and those obtained by the models are, in general, subtle. However, when the whole volume is considered, small differences involving the voxels of this label translate into a large difference in IoU. Such mislabeling for the degradation layer was observed for almost the entire predicted segmentation volumes for those models even though normalization of the SRμCT data was performed to prevent poor generalization performance. By normalizing the voxel values, we bring them to a common scale, enabling the network to focus on more relevant information. Without normalization, a CNN might assign different weights to similar features due to variations in contrast, resulting in decreased generalization performance. Moreover, it is also possible to observe that in certain instances, the obtained predictions seem to have a more accurate segmentation than the segmentation proposed by ground truth data. This is especially true for the U-Net and the 2D and 3D nnU-Net. Example slices for sample 2 can be seen in Supplementary Appendix Figure SA2.

The predicted data were also quantitatively analyzed against the reference data by comparing degradation and osseointegration parameters such as the DR, BIC, and BV/TV. Figure 4 shows the mean relative error calculated between the above-mentioned parameters, considering the ground truth and predicted segmentation

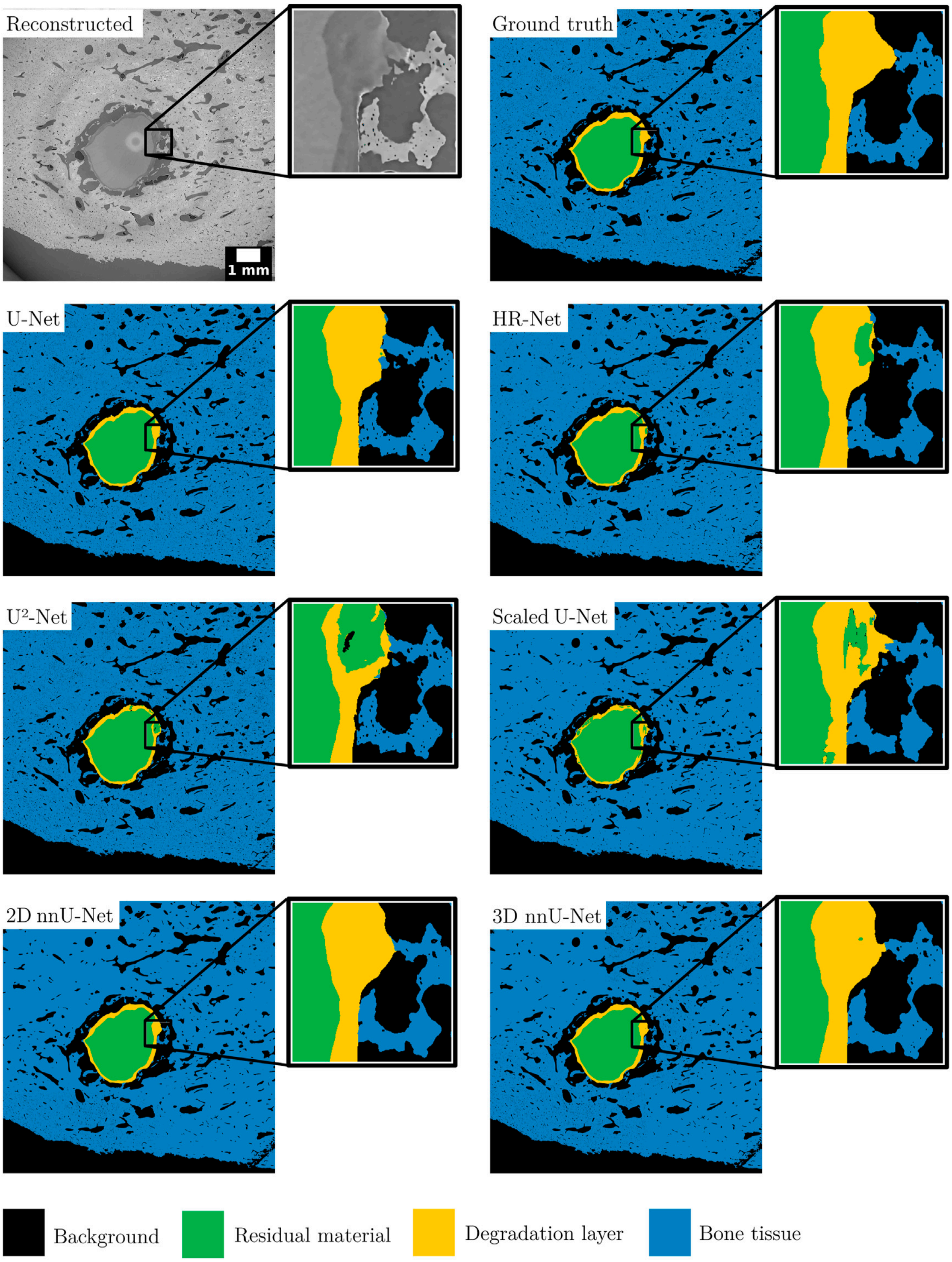

FIGURE 3
Representative slices of image data from one test sample and the corresponding predicted segmented data sets for all tested models. The bright circle in the middle of the reconstructed and denoised image is an artifact from the image stitching process [11]. All the insets correspond to a zoom-in area of 1 mm^2.

of both tested samples. The detailed results can be found in Table 5. Overall, the relative error for the DR and BV/TV parameters showed less deviation from the ground truth values (less than 5%) for all tested models. In contrast, the BIC parameter differed significantly, achieving mean relative errors of more than 120%. The DR parameter is calculated based on the initial and final implant volumes (Eq. 4), where the latter is dependent on the residual material label. From this perspective, the lower deviation from the ground truth DR values is in agreement with the high

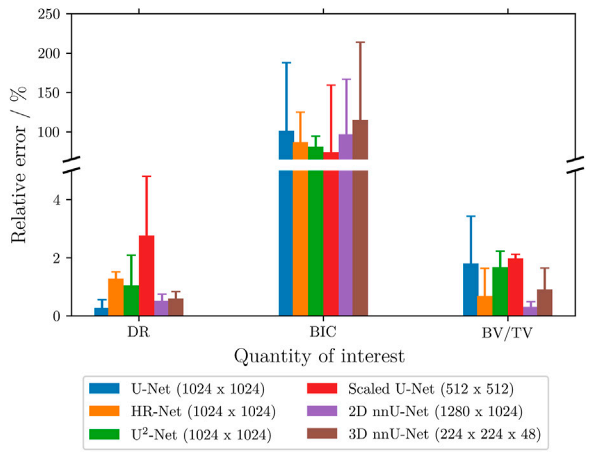

FIGURE 4
Relative error for degradation and osseointegration parameters. The segmentations predicted with the two test samples were used to calculate the parameters for implant degradation and osseointegration. The graph shows the mean value of the relative error for the two samples compared to the reference segmentation and each bar shows the standard deviation (error bars).

IoU values obtained for this label for both samples. Regarding the BIC parameter, the values obtained with the predicted segmentations showed to be higher than the reference ones. The BIC is strongly related to the degradation layer, and therefore it is very sensitive to mislabeling. In fact, for sample 1, this error becomes smaller because the number of contact voxels between the degradation layer and the bone tissue is also less due to the morphology of the implant and its surroundings. However, this error becomes much more expressive in sample 2, as there is much more bone tissue in contact with the degradation layer considering the whole volume analyzed. Finally, the parameter BV/TV, which is calculated considering a 1 mm ROI, is only dependent on the bone tissue and background. As stated earlier, those labels achieved high IoU for all tested models, thus explaining the negligible error below 2% for the majority of models and samples.

3.2 Prediction of unlabeled data

The 2D nnU-Net was used to predict the segmentation of unlabeled SRμCT data. This data refers to ZX00 implants explanted from sheep at different healing times (6, 12, and 24 weeks). The prediction of samples was performed for three samples for each time point. Figure 5 shows the denoised slice after tomographic reconstruction and the predicted segmentation by the 2D nnU-Net, for a selected cross-section for one sample per healing time considered. In addition, the insets display a magnification of problematic points of the segmentation of the deep learning prediction. It is important to point out that the samples chosen were images that could exemplify most of the problems encountered in the predictions of the segmented volumes. For the 12-week sample, it is possible to see that some voxels from the residual metal label were mislabeled into degradation layer voxels. In addition, the bone tissue label also needed to be corrected, because the CNN failed to predict the bone lacunae, which should have been assigned to the background label. In fact, this problem was found in all samples analyzed, indicating that the CNN model had difficulties identifying small features. This happened even though these small features were segmented as background within the ground truth segmentations used for the CNN training. Moreover, in the 24-week sample, it is possible to see that there was mislabeling of the background label in the region corresponding to the residual material label. Although the datasets used for prediction are from a different source than the datasets used for training/testing and validation, image quality metrics such as SNR and CNR were calculated to confirm that the image quality, when considering key parameters, remains consistent. Additionally, previous research has highlighted the effectiveness of CNN models trained using diverse data sources. This indicates that despite inherent differences among data sources, CNN models showcase adaptability across varied imaging conditions [62].

Due to the incorrect segmentation in parts, (semi-)manual corrections become necessary before quantification of the degradation rate and other parameters is possible. Figure 5 is also showing the corrected segmentation. Despite the time investment in training CNNs, their integration streamlines segmentation, allowing faster processing even with manual adjustments. Tools utilizing CNN-generated outputs as a foundation expedite manual corrections, notably accelerating the segmentation of extensive synchrotron data. While not eliminating manual segmentation entirely, this hybrid approach significantly expedites the process, offering a time-efficient solution while upholding segmentation quality [10, 32]. To further enhance the generalizability of the model it should, therefore, be updated continuously, e.g., in a loop, using a small

TABLE 5 Degradation and osseointegration metrics from ground truth and predicted segmentations. Ground truth values are in bold. The values in the brackets are the relative errors for each measured parameter against the ground value. (–) represents the errors that are less than 1%.

Parameter	Segmentations							
	Sample	G. Truth	U-Net	HR-Net	U^2-Net	Scaled U-Net	2D nnU-Net	3D nnU-Net
DR/mm · a^{-1}	1	**0.58**	0.58 (–)	0.57 (−1%)	0.58 (–)	0.57(-1%)	0.58 (–)	0.58 (–)
	2	**0.36**	0.36 (–)	0.35 (−1%)	0.36 (–)	0.34 (−4%)	0.36 (–)	0.36 (–)
BIC/%	1	**5.25**	7.36 (+40%)	8.42 (+60%)	9.02 (+71%)	5.99 (+14%)	7.76 (+47%)	7.62 (+45%)
	2	**7.48**	19.65 (+162%)	16.01 (+114%)	14.25 (+90%)	17.54 (+134%)	18.44 (+146%)	21.31 (+184%)
BV/TV/%	1	**22.48**	21.81 (−2%)	22.18 (−1%)	22.02 (−2%)	22.02 (−2%)	22.44 (–)	22.40 (–)
	2	**56.83**	55.60 (–)	57.01 (–)	55.70 (−1%)	57.90 (+1%)	57.08 (–)	57.65 (+1%)

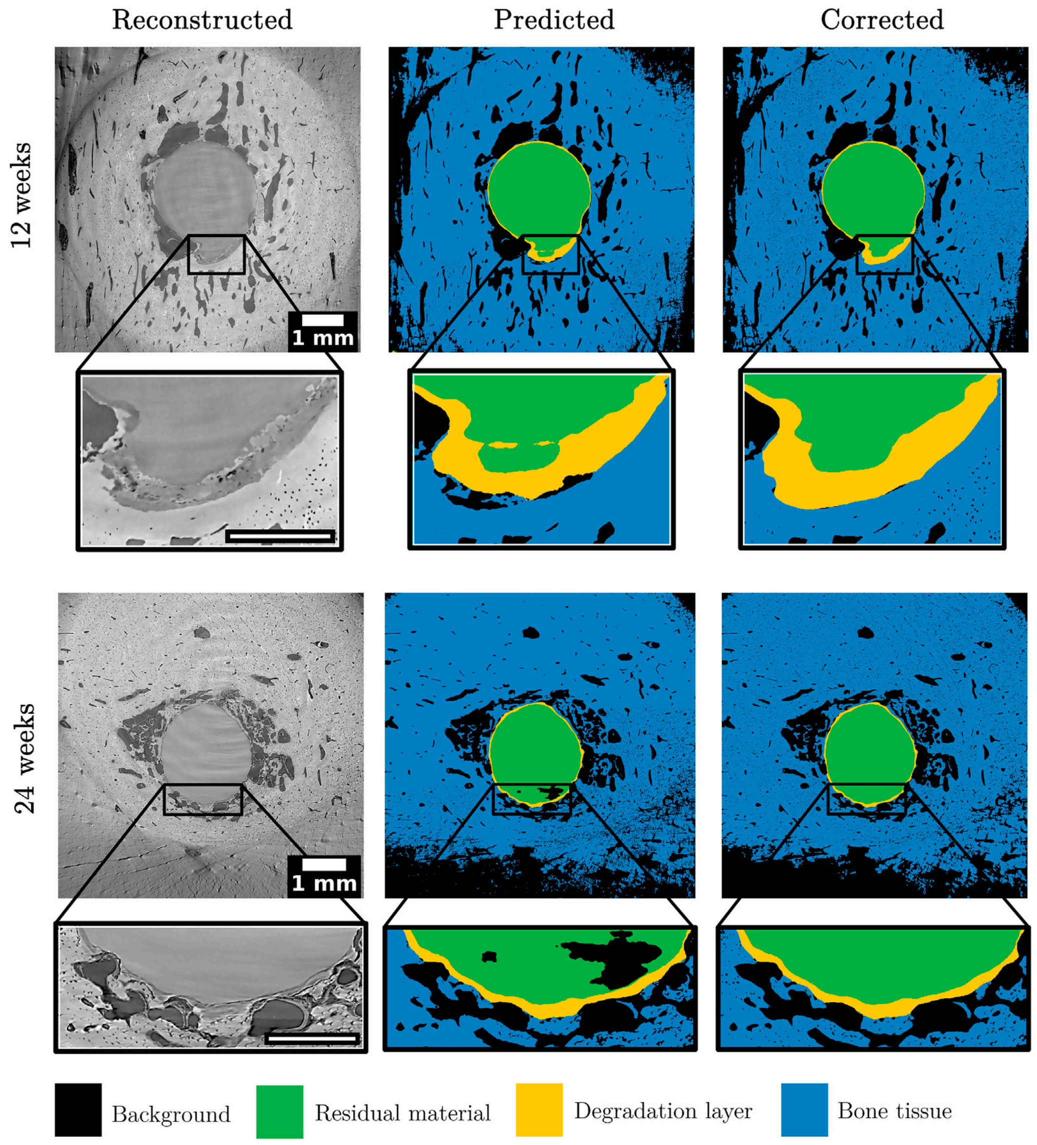

FIGURE 5
Selected cross-sections of the reconstructed and denoised images, along with their predicted and corrected segmentations, for one sample per healing time considered. The images also display insets where it is possible to compare problematic results from the predicted segmentations and the corrections which were done in order to calculate the degradation and osseointegration parameters with reliability. The scale of the insets is 0.5 mm.

number of corrected slices and various SRμCT datasets. This could be done by an active learning approach [52, 63]. By doing so and adding model-in-the-loop annotations [64], one may ultimately arrive at a more powerful tool for the semantic segmentation of SRμCT data. In addition, the frameworks used in this work could take advantage of recent studies that seek to define the key problems that lead algorithms to fail in tasks such as semantic segmentation, in order to obtain more robust CNN models for this challenging task [65, 66].

4 Conclusion

This work aimed to study the use of deep learning for the semantic segmentation of SRμCT data of biodegradable Mg-based implants. Three CNN frameworks were used for comparison and their performance in terms of mIoU was analyzed. It became clear in this evaluation that the image input size influences the mIoU. Among the tested and validated models, the best-performing ones were the nnU-Net 2D and 3D, which obtained mIoU performances greater than 93%, while the others had mIoU less than 91%. However, identifying the degradation layer label proved challenging due to its morphological heterogeneity and low contrast in the input images. Quantitative evaluation of the segmentations revealed smaller errors for degradation and osseointegration parameters, except for bone-implant contact (BIC), where high errors were observed. The limitations of the CNNs in identifying the degradation layer were considered the main cause of the higher errors. In conclusion, the deep learning architectures proved to be capable of obtaining a good performance in the semantic segmentation of SRμCT data but could be further improved by continuous retraining.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The animal study was approved by the Austrian Federal Ministry for Science and Research. The work adhered to the guidelines outlined by the European Convention for the Protection of Vertebrate Animals Used for Experimental and Other Scientific Purposes. The study was conducted in accordance with the local legislation and institutional requirements.

Author contributions

AL: Formal Analysis, Investigation, Software, Visualization, Writing–original draft. BK: Resources, Software, Writing–review and editing. HĆ: Formal Analysis, Investigation, Supervision, Writing–review and editing. RM: Data curation, Investigation, Writing–review and editing. FB: Data curation, Investigation, Software, Writing–review and editing. RW-R: Resources, Supervision, Writing–review and editing. JM: Funding acquisition, Software, Supervision, Writing–review and editing. BZ-P: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors acknowledge funding via the projects “Multi-task Deep Learning for Large-scale Multimodal Biomedical Image Analysis” (MDLMA, BMBF project number 031L0202A), the Helmholtz AI (HAICU) project Universal Segmentation Framework (UniSeF). This project has also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 811226 (MgSafe).

Acknowledgments

We acknowledge the allocation of beamtime at beamline P05 and P07 at PETRA III at the Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany, granted for proposals 20200278 (beamtimes 11009431) and 20191467 (beamtime 11008741). This research was supported in part through the Maxwell computational resources operated at Deutsches Elektronen-Synchrotron, Hamburg, Germany. The authors thank RM for performing the animal experiments.

Conflict of interest

Authors AL, BK, HĆ, FB, RW-R, JM, and BZ-P were employed by Helmholtz-Zentrum Hereon GmbH. Author RM was employed by the Medical University of Graz.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2024.1257512/full#supplementary-material

References

1. Larrue A, Rattner A, Peter ZA, Olivier C, Laroche N, Vico L, et al. Synchrotron radiation micro-CT at the micrometer scale for the analysis of the three-dimensional morphology of microcracks in human trabecular bone. *PLOS ONE* (2011) 6(7): e21297–12. doi:10.1371/journal.pone.0021297

2. Swolfs Y, Morton H, Scott A, Gorbatikh L, Reed P, Sinclair I, et al. Synchrotron radiation computed tomography for experimental validation of a tensile strength model for unidirectional fibre-reinforced composites. *Composites A: Appl Sci Manufacturing* (2015) 77:106–13. doi:10.1016/j.compositesa.2015.06.018

3. Luo Y, Wu SC, Hu YN, Fu YN. Cracking evolution behaviors of lightweight materials based on *in situ* synchrotron X-ray tomography: a review. *Front Mech Eng* (2018) 13(4):461–81. doi:10.1007/s11465-018-0481-2

4. Zeller-Plumhoff B, Tolnai D, Wolff M, Greving I, Hort N, Willumeit-Römer R. Utilizing synchrotron radiation for the characterization of biodegradable magnesium alloys—from alloy development to the application as implant material. *Adv Eng Mater* (2021) 23(11):2100197. doi:10.1002/adem.202100197

5. Seitz J-M, Lucas A, Kirschner M. Magnesium-based compression screws: a novelty in the clinical use of implants. *JOM* (2016) 68(4):1177–82. doi:10.1007/s11837-015-1773-1

6. Bowen PK, Shearier ER, Zhao S, Guillory RJ, Zhao F, Goldman J, et al. Biodegradable metals for cardiovascular stents: from clinical concerns to recent Zn-alloys. *Adv Healthc Mater* (2016) 5(10):1121–40. doi:10.1002/adhm.201501019

7. Kačarević ŽP, Rider P, Elad A, Tadic D, Rothamel D, Sauer G, et al. Biodegradable magnesium fixation screw for barrier membranes used in guided bone regeneration. *Bioactive Mater* (2022) 14:15–30. doi:10.1016/j.bioactmat.2021.10.036

8. Galli S. *On magnesium-containing implants for bone applications*. Malmö University, Faculty of Odontology OD (2016). Doctoral Dissertation in Odontology.

9. Krüger D, Zeller-Plumhoff B, Wiese B, Yi S, Zuber M, Wieland DF, et al. Assessing the microstructure and *in vitro* degradation behavior of Mg-xGd screw implants using µCT. *J Magnesium Alloys* (2021) 9(6):2207–22. doi:10.1016/j.jma.2021.07.029

10. Krüger D, Galli S, Zeller-Plumhoff B, Wieland DF, Peruzzi N, Wiese B, et al. High-resolution *ex vivo* analysis of the degradation and osseointegration of Mg-xGd implant screws in 3D. *Bioactive Mater* (2022) 13:37–52. doi:10.1016/j.bioactmat.2021.10.041

11. Marek R, Ćwieka H, Donohue N, Holweg P, Moosmann J, Beckmann F, et al. Degradation behavior and osseointegration of Mg–Zn–Ca screws in different bone regions of growing sheep: a pilot study. *Regenerative Biomater* (2022) 10:rbac077. doi:10.1093/rb/rbac077

12. Sefa S, Wieland DF, Helmholz H, Zeller-Plumhoff B, Wennerberg A, Moosmann J, et al. Assessing the long-term *in vivo* degradation behavior of magnesium alloys - a high resolution synchrotron radiation micro computed tomography study. *Front Biomater Sci* (2022) 1. doi:10.3389/fbiom.2022.925471

13. Wang Y, Miller JD. Current developments and applications of micro-CT for the 3D analysis of multiphase mineral systems in geometallurgy. *Earth-Science Rev* (2020) 211:103406. doi:10.1016/j.earscirev.2020.103406

14. Withers PJ, Bouman C, Carmignato S, Cnudde V, Grimaldi D, Hagen CK, et al. X-ray computed tomography. *Nat Rev Methods Primers* (2021) 1(1):18. doi:10.1038/s43586-021-00015-4

15. Galli S, Hammel JU, Herzen J, Damm T, Jimbo R, Beckmann F, et al. Evaluation of the degradation behavior of resorbable metal implants for *in vivo* osteosynthesis by synchrotron radiation based x-ray tomography and histology. *SPIE Proc* (2016) 9967: 996704. doi:10.1117/12.2237563

16. Moosmann J, Wieland DCF, Zeller-Plumhoff B, Galli S, Krüger D, Ershov A, et al. A load frame for *in situ* tomography at PETRA III. In: Müller B, Wang G, editors. *Proc. SPIE 11113, developments in X-ray tomography XII*. San Diego, California, United States (2019). doi:10.1117/12.2530445

17. Bockelmann N, Diana Krüger DC, Wieland F, Zeller-Plumhoff B, Peruzzi N, Galli S, et al. Sparse annotations with random walks for U-net segmentation of biodegradable bone implants in synchrotron microtomograms. In: International Conference on Medical Imaging with Deep Learning (MIDL 2019 – Extended Abstract Track) (2019). doi:10.48550/arXiv.1908.04173

18. Menze R, Hesse B, Kusmierczuk M, Chen D, Weitkamp T, Bettink S, et al. Synchrotron microtomography reveals insights into the degradation kinetics of bio-degradable coronary magnesium scaffolds. *Bioactive Mater* (2024) 32:1–11. doi:10.1016/j.bioactmat.2023.09.008

19. Hao S, Zhou Y, Guo Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing* (2020) 406:302–21. doi:10.1016/j.neucom.2019.11.118

20. Furat O, Wang M, Neumann M, Petrich L, Weber M, Krill CE, et al. Machine learning techniques for the segmentation of tomographic image data of functional materials. *Front Mater* (2019) 6:2296–8016. doi:10.3389/fmats.2019.00145

21. Varfolomeev I, Yakimchuk I, Safonov I. An application of deep neural networks for segmentation of microtomographic images of rock samples. *Computers* (2019) 8(4): 72. doi:10.3390/computers8040072

22. Malimban J, Lathouwers D, Qian H, Verhaegen F, Wiedemann J, Brandenburg S, et al. Deep learning-based segmentation of the thorax in mouse micro-CT scans. *Scientific Rep* (2022) 12(1):1822. doi:10.1038/s41598-022-05868-7

23. Ajit A, Acharya K, Samanta A. A review of convolutional neural networks. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE); Vellore, India (2020). p. 1–5. doi:10.1109/ic-ETITE47903.2020.049

24. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* (2021) 8(1):53. doi:10.1186/s40537-021-00444-8

25. Ulku I, Akagündüz E. A survey on deep learning-based architectures for semantic segmentation on 2D images. *Appl Artif Intelligence* (2022) 36(1):2032924. doi:10.1080/08839514.2022.2032924

26. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. *Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science*, 9351. Springer, Cham (2015). p. 234–41. doi:10.1007/978-3-319-24574-4_28

27. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015). p. 3431–40. doi:10.1109/CVPR.2015.7298965

28. Chlebus G, Schenk A, Moltz JH, van Ginneken B, Hahn HK, Meine H. Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing. *Scientific Rep* (2018) 8(1):15497. doi:10.1038/s41598-018-33860-7

29. Kohl S, Romera-Paredes B, Meyer C, Fauw JD, Ledsam JR, Maier-Hein KH, et al. A probabilistic U-net for segmentation of ambiguous images. In: *Advances in neural information processing systems*. NeurIPS (2018). p. 31. doi:10.48550/arXiv.1806.05034

30. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. *Medical image segmentation review: the success of U-net* (2022). doi:10.48550/arXiv.2211.14830

31. Yin T-K, Wu L-Y, Hong T-P. Axial attention inside a U-net for semantic segmentation of 3D sparse LiDAR point clouds. In: *2022 IEEE intelligent vehicles Symposium (IV)* (2022). p. 1543–9. doi:10.1109/IV51971.2022.9827257

32. Baltruschat IM, Ćwieka H, Krüger D, Zeller-Plumhoff B, Schlünzen F, Willumeit-Römer R, et al. Scaling the U-net: segmentation of biodegradable bone implants in high-resolution synchrotron radiation microtomograms. *Scientific Rep* (2021) 11(1):24237. doi:10.1038/s41598-021-03542-y

33. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* (2021) 18(2):203–11. doi:10.1038/s41592-020-01008-z

34. Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The medical segmentation Decathlon. *Nat Commun* (2022) 13(1):4128. doi:10.1038/s41467-022-30695-9

35. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognition* (2020) 106:107404. doi:10.1016/j.patcog.2020.107404

36. ran Wang R, Wang Y. Improved U2net-based liver segmentation. In: Proceedings of the 5th International Conference on Advances in Image Processing. ICAIP '21 (2022). p. 48–55. doi:10.1145/3502827.3502832

37. Shao J, Zhou K, Cai YH, Geng DY. Application of an improved U2-net model in ultrasound median neural image segmentation. *Ultrasound Med Biol* (2022) 48(12): 2512–20. doi:10.1016/j.ultrasmedbio.2022.08.003

38. Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). doi:10.48550/arXiv.1902.09212

39. Gong K, Liang X, Zhang D, Shen X, Lin L. *Look into person: self-supervised structure-sensitive learning and A new benchmark for human parsing* (2017). arXiv: 1703.05446. doi:10.48550/arXiv.1703.05446

40. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. *The cityscapes dataset for semantic urban scene understanding* (2016). arXiv: 1604.01685. doi:10.48550/arXiv.1604.01685

41. Kazimi B. TomoSeg: a semantic segmentation framework for tomography data (2022). Available at: https://gitlab.desy.de/bashir.kazimi/tomoseg.

42. Nowozin S. Optimal decisions from probabilistic models: the intersection-over-union case. In: IEEE Conference on Computer Vision and Pattern Recognition (2014). doi:10.1109/CVPR.2014.77

43. Holweg P, Berger L, Cihova M, Donohue N, Clement B, Schwarze U, et al. A lean magnesium–zinc–calcium alloy ZX00 used for bone fracture stabilization in a large growing-animal model. *Acta Biomater* (2020) 113:646–59. doi:10.1016/j.actbio.2020.06.013

44. Wilde F, Ogurreck M, Greving I, Hammel JU, Beckmann F, Hipp A, et al. Micro-CT at the imaging beamline P05 at PETRA III. In: AIP Conference Proceedings 1741 (2016). doi:10.1063/1.4952858

45. Schell N, Martins RV, Beckmann F, Ruhnau HU, Kiehn R, Schreyer A. The high energy materials science beamline at PETRA III. In: *Materials science forum - mater SCI forum* (2008). p. 571–2. doi:10.4028/www.scientific.net/MSF.571-572.261

46. Moosmann J, Ershov A, Weinhardt V, Baumbach T, Prasad MS, LaBonne C, et al. Time-lapse X-ray phase-contrast microtomography for *in vivo* imaging and analysis of morphogenesis. *Nat Protoc* (2014) 9:294–304. doi:10.1038/nprot.2014.033

47. Moosmann J. *moosmann/matlab*: *version v1.0* (2021). doi:10.5281/zenodo.5118737

48. Palenstijn WJ, Batenburg KJ, Sijbers J. Performance improvements for iterative electron tomography reconstruction using graphics processing units (GPUs). *J Struct Biol* (2011) 176(2):250–3. doi:10.1016/j.jsb.2011.07.017

49. van Aarle V, Palenstijn WJ, De Beenhouwer J, Altantzis T, Bals S, Batenburg KJ, et al. The ASTRA Toolbox: a platform for advanced algorithm development in electron tomography. *Ultramicroscopy* (2015) 157:35–47. doi:10.1016/j.ultramic.2015.05.002

50. Bruns S, Stipp SLS, Sørensen HO. Looking for the Signal: a guide to iterative noise and artefact removal in X-ray tomographic reconstructions of porous geomaterials. *Adv Water Resour* (2017) 105:96–107. doi:10.1016/j.advwatres.2017.04.020

51. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction* 2nd ed. New York: Springer Series in Statistics. Springer (2009). doi:10.1007/978-0-387-84858-7

52. Kazimi B, Heuser P, Schluenzen F, Cwieka H, Krüger D, Zeller-Plumhoff B, et al. An active learning approach for the interactive and guided segmentation of tomography data. In: Müller B, Wang G, editors. *Developments in X-ray tomography XIV. Society of photo-optical instrumentation engineers (SPIE) conference series 12242* (2022). p. 122420F. doi:10.1117/12.2637973

53. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. *PeerJ* (2014) 2:e453. doi:10.7717/peerj.453

54. Gonzalez J, Hou RQ, Nidadavolu EP, Willumeit-Römer R, Feyerabend F. Magnesium degradation under physiological conditions – best practice. *Bioactive Mater* (2018) 3(2):174–85. doi:10.1016/j.bioactmat.2018.01.003

55. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature* (2020) 585(7825):357–62. doi:10.1038/s41586-020-2649-2

56. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* (2020) 17(3):261–72. doi:10.1038/s41592-019-0686-2

57. Ulyanov D, Vedaldi A, Lempitsky V. *Instance normalization: the missing ingredient for fast stylization* (2017). ArXiv. doi:10.48550/arXiv.1607.08022

58. Kolarik M, Burget R, Riha K. Comparing normalization methods for limited batch size segmentation neural networks. In: 43rd International Conference on Telecommunications and Signal Processing (TSP) (2020). p. 677–80. doi:10.1109/TSP49548.2020.9163397

59. Andrew AL, Maas L. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on Machine Learning, 28 (2013). p. 3.

60. Parico AI, Ahamed T. Real time pear fruit detection and counting using YOLOv4 models and deep SORT. *Sensors* (2021) 21:4803. doi:10.3390/s21144803

61. Szabó A, Jamali-Rad H, Mannava S-D. Tilted cross-entropy (TCE): promoting fairness in semantic segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 19-25 June 2021; Nashville, TN, USA. IEEE (2021). p. 2305–10. doi:10.1109/CVPRW53098.2021.00261

62. Aust O, Thies M, Weidner D, Wagner F, Pechmann S, Mill L, et al. Tibia cortical bone segmentation in micro-CT and X-ray microscopy data using a single neural network. In: Maier-Hein K, Deserno TM, Handels H, Maier A, Palm C, Tolxdorff T, editors. *Bildverarbeitung für die Medizin 2022. Informatik aktuell*. Springer Vieweg, Wiesbaden (2022). p. 333–8. doi:10.1007/978-3-658-36932-3_68

63. Michen M, Haßler U. Deep learning and active learning based semantic segmentation of 3D CT data. In: LÃngle T, Heizmann M, editors. *Forum bildverarbeitung 2022 image processing forum 2022* (2022). p. 163–73. doi:10.5445/KSP/1000150865

64. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. *Segment anything* (2023). arXiv:2304.02643. doi:10.48550/arXiv.2304.02643

65. Reinke A, Grab G, Maier-Hein L. Challenge results are not reproducible. In: Deserno TM, Handels H, Maier A, Maier-Hein K, Palm C, Tolxdorff T, editors. *Bildverarbeitung für die Medizin 2023. BVM 2023. Informatik aktuell.* Springer Vieweg, Wiesbaden (2023). p. 198–203. doi:10.1007/978-3-658-41657-7_43

66. Roß T, Bruno P, Reinke A, Wiesenfarth M, Koeppel L, Full PM, et al. Beyond rankings: learning (more) from algorithm validation. *Med Image Anal* (2023) 86:102765. doi:10.1016/j.media.2023.102765