Evaluation of deeply supervised neural networks for 3D pore segmentation in additive manufacturing

D. Iuso[a,e], S. Chatterjee[b], R. Heylen[c], S. Cornelissen[d], J. De Beenhouwer[a,e], and J. Sijbers[a,e]

[a]imec-Vision Lab, Departement of Physics, University of Antwerp, Belgium
[b]Faculty of Computer Science, Otto von Guericke University, Magdeburg, Germany
[c]Flanders Make vzw., Gaston Geenslaan 8, 3001 Leuven, Belgium
[d]Materialise NV., Technologielaan 15, 3001 Leuven, Belgium
[e]DynXlab: Center for 4D Quantitative X-ray Imaging and Analysis, Antwerp, Belgium

ABSTRACT

Additive manufacturing (AM) is increasingly gaining interest as a low-waste production technique, capable of producing objects using a computer-aided design file. It is particularly interesting for rapid prototyping of parts and manufacturing objects that have complex shapes. However, as in the case of AM through selective laser melting (SLM), manufactured objects may contain defects that can seriously alter their properties. These defects may appear as pores, which can be detected by X-ray computed tomography (X-CT) in a non-destructive manner. CT images can simply be segmented by thresholding or through more advanced techniques such as discrete X-CT reconstruction or machine learning techniques. Nevertheless, these techniques are vulnerable to image reconstruction artefacts. In this work, we evaluate the performance of state-of-the-art, deeply supervised 3D deep learning networks (UNet++, UNet 3+ and UNet-MSS) in terms of segmentation performance of pores from X-ray CT images. The networks have been trained on a real CT dataset, with (noisy) labels produced from both conventional thresholding of the CT images as well as more advanced discrete polychromatic reconstructions. Furthermore, the performance of the networks was evaluated on a test dataset with severe CT artefacts. Pore segmentation from real CT images, which include noise and reconstruction artefacts, revealed that the best performing network was UNet++ with an average Sørensen–Dice score of 0.869 ± 0.006.

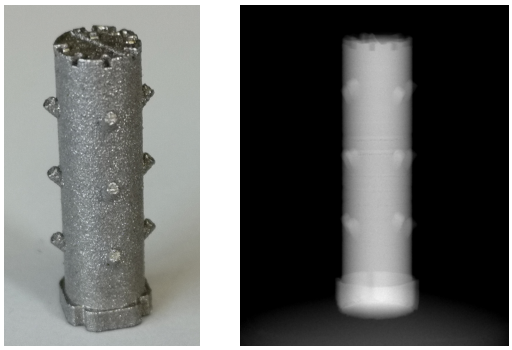
Keywords: Additive manufacturing, anomaly detection, pore segmentation, X-ray CT, 3D patch-based segmentation, selective laser melting

1. INTRODUCTION

Through additive manufacturing, it is possible to manufacture objects layer by layer with a wide range of materials and shapes, which is appealing for fields that span from aerospace to orthodontics. More specifically, metal objects can be manufactured through the SLM process, although this printing technique is known to create defects within the manufactured sample due to a non-ideal printing process.[1] For both the printing process evaluation and the quality assurance, a non-destructive inspection of the entire sample is required, which is commonly performed through X-ray computed tomography (X-CT). Elements of interest during the (non-destructive) analysis of these samples include the internal and external surface properties,[2] the structural integrity of samples,[3] as well as the identification and quantification of the number of defects that arise from the AM process.[4]

Defects can be detected by processing the X-CT images with conventional methods,[5–7] deep learning[8] and potentially through discrete algebraic X-ray reconstruction techniques (e.g., poly-DART[9]). The multiple sources of X-CT image artefacts (e.g., beam hardening, cone-beam artefacts, under-sampling, scattered radiation) and noise in the acquired X-ray projections severely hinder the performance of threshold methods.[5] To reduce the number of voxels that are wrongly classified as pores, a rule of thumb is to identify as defects only voxels whose gray level substantially deviates from the mean attenuation value in an 8-connected 3D neighbourhood in the X-CT volume.[6] To counteract these detrimental effects, deep learning techniques are lately being employed

E-mail: Domenico.Iuso@uantwerpen.be



(a) (b)

Figure 1: On the left, a picture of one sample used for training. On the right, the related X-ray radiograph.

for defect segmentation in X-CT images of AM samples, although most of the recent literature is focused on detecting defects in a 2D image. Reducing the complexity of the analysis from a 3D volume to a series of 2D images circumvents the hardware limitations and reduces the training dataset, although the networks do not fully exploit the volumetric spatial information. For a more reliable pore segmentation with robustness against image artefacts, a possible solution to the above-mentioned problems is to use a 3D patch-based approach. Moreover, this approach allows the networks to be applied to the analysis and classification of samples of any physical dimension.

In this study, we investigate how the information from an Otsu thresholding-based and a discrete polychromatic reconstruction can be used to train recent 3D convolutional neural networks for the pore segmentation task. The four deep learning models used for this study are: UNet,[10] UNet-MSS,[11] UNet++[12] and UNet 3+.[13] While the UNet model has been trained as a baseline, the last three are deeply supervised networks that are supposed to yield more reliable results, since the hidden layers of the networks are enticed to comply with the desired output. After training on AM samples of different materials and printing procedures using a 3D patch-based approach, the networks could be used to identify porosity even when X-CT artefacts are present.

2. METHODS

A customized Tescan UnitomXL scanner (FleXCT)[14] was used to scan a total of four AM samples, for which a detailed description is provided in section 2.1 along with an explanation of their acquisition setup and X-CT reconstruction. The X-CT images and the related labels were used to train the networks with a 3D patch-based approach, where data augmentation was used to enrich the training-sets. How the networks were built and how the trainings were performed is described in sections 2.2 and 2.3. Eventually, the performance of the networks were evaluated on the validation dataset and on a testing dataset with severe artefacts.

2.1 Dataset

Three cylinders were printed from stainless steel 316L, CoCr alloy and TiAl6V4 powder; their diameter was 0.5 mm and height 20 mm, with ∼4% tolerance due to the printing process, cutting and polishing procedure. Each of the samples was acquired for a total of 4283 projections with an source-to-detector distance of 650 mm and source-to-object distance of 43.333 mm. The X-ray source was operated with 230 kVp for the first sample with a 1 mm Cu filter, 220 kVp for the second sample with a 1.5 mm Cu filter and 210 kVp for the third sample with a 1 mm Cu filter. All the above-mentioned samples were printed with off-nominal laser parameters in order to induce porosity. An picture of a cylindrical sample a related X-ray projection is shown in Fig.1. The X-ray projections of each sample were reconstructed with the FDK[15] and polychromatic SIRT (abbreviated with poly-SIRT, following the implementation described in[9]) algorithms and were, respectively, labelled with Otsu and poly-DART. The FDK and a proprietary beam-heardening correction are part of the FleXCT software, while poly-SIRT and poly-DART where implemented on the ASTRA toolbox.[16] For all the samples, the FDK and polySIRT reconstructions have the same 10 µm voxel-size, have approximately 1 billion voxels and were pre- and post-processed to account for beam-hardening, ring artefacts, scattered radiation.[17]

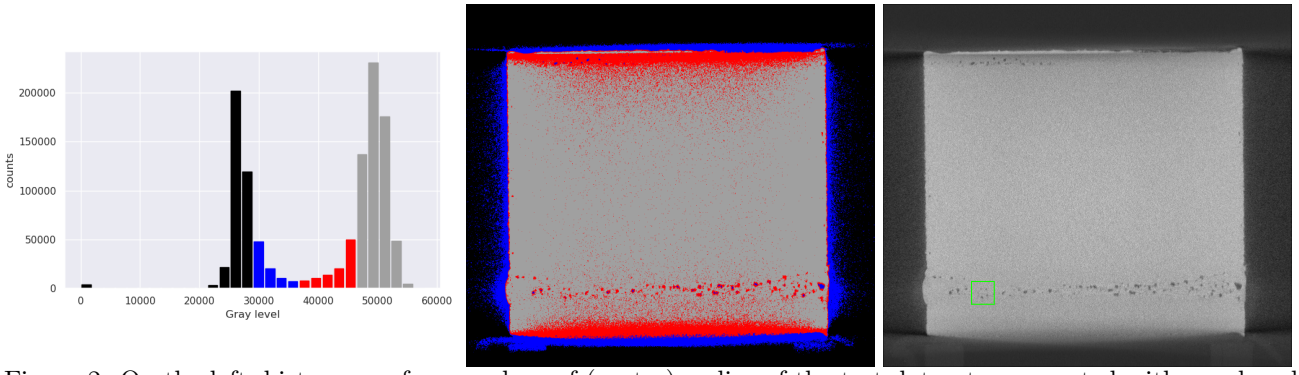


Figure 2: On the left, histogram of grey values of (centre) a slice of the test dataset represented with a reduced number of colours, In absence of artefacts, the black colour indicates that in a voxel there is no material, while the red and blue voxels indicate pores that are partially-filled with the material of the sample. Due to artefacts, threshold-based methods for pore-segmentation would create many voxel-wise misclassifications. On the right, the same slice with gray values linearly linked to the attenuation values. A region of interest (green) is further discussed in Fig. 6.

The segmentation techniques produced binary masks as shown in Fig. 3. Three training-sets were built with the X-CT reconstruction of the first two samples in order to assess the influence of the labelling technique on the performance of the networks. The first training-set was composed by the FDK reconstructions of the first two samples labelled with Otsu, the second training-set by their poly-SIRT reconstruction labelled with poly-DART, and the third training-set by the union of the above-mentioned training-sets. The validation-set is composed by the FDK and poly-SIRT reconstructions of the third sample, labelled with both segmentation techniques. Noise and image artefacts may influence the segmentation results and produce incorrect labels. Nevertheless, the impact of the noisy labels during training is diminished by the use of data augmentation,[18] which has been used also to expand the training-sets. The techniques of the data augmentation are shown in Table 1 and aim to create new spatial configurations (flip of patches in random directions and elastic distortion) and to teach the networks to be robust against noise, the specific attenuation of samples and cone-beam/undersampling/beam-hardening artefacts.

	Technique	Typology	Likelihood
One of	3D Random flip	Spatial	0.5
	3D Elastic distortion	Spatial	0.5
One of	Gaussian noise	Intensity	0.4
	Random re-scaling	Intensity	0.4
	Random Fourier wave	Intensity	0.2

Table 1: Data augmentation techniques used during trainings at each epoch.

To test the performance of the networks on a CT reconstruction with severe artefacts and when the object has a different shape, an additional (fourth) sample was used. It was a stainless steel 316L cube with ad edge of 10 mm. Its FDK reconstruction (Fig. 2) suffered from many artefacts such as scatter, cone-beam artefacts, residual beam-hardening artefacts, etc. As a result, pores appeared with lower contrast than in reconstructions of the other samples, which makes the segmentation of pores more challenging.

2.2 Networks

Since deeply supervised learning has shown great performances in a variety of segmentation tasks,[11–13] in this study UNet-MSS, UNet++ and UNet 3+ have been trained and compared, along with the traditional UNet. The networks have a comparable number of parameters (Unet: 1.4 M, UNet-MSS: 1.4 M, UNet++: 1.6 M, UNet 3+: 1.7 M) and the convolutional layers of the decoder and encoder paths have the same morphology: Double

(a) (b) (c) (d) (e)

Figure 3: In (a) is shown a slice of the polySIRT reconstruction of the validation sample; (b) A 150x150 pixel region of interest; (c) polyDART mask; (d) Otsu segmentation and (e) prediction of the network that shown the best performance on this dataset.

convolution (each followed by a LeakyRelu activation), Dropout (p=0.1). The convolutional layers' parameters were initialised with Kaiming method,[19] since they are followed by a non-symmetric activation function, with the exception of the output layers that were initialised with Xavier's method because of the sigmoid activation function.[20]

2.3 Training

The training of all the networks was achieved by minimising the Focal Tversky loss:[21]

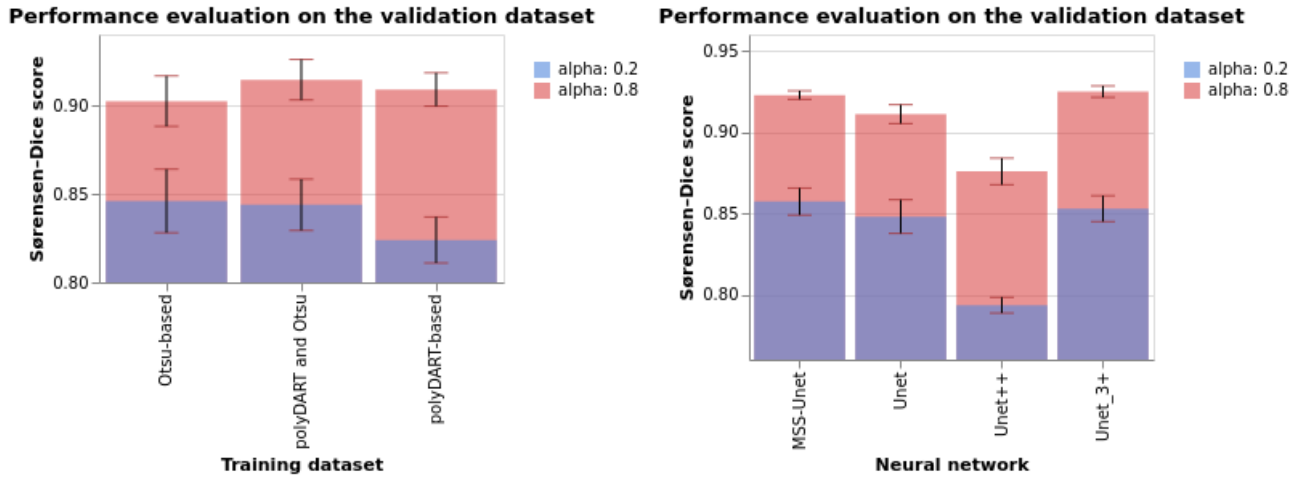
$$FTL = \left(1 - \frac{TN}{TN + \alpha FN + \beta FP}\right)^{\gamma} \tag{1}$$

In the case of deeply supervised networks, the loss is computed as the sum of the FTL of all the supervised layers. The use of this particular loss function is motivated by the class imbalance that is present between the frequency of positive labels (namely, voxels that contain material) and negative labels, which is a degree of freedom that the traditional Sørensen–Dice loss does not deliver.[8] The parameter γ was set to 0.5 in order to accelerate the convergence to a meaningful solution, since it better rewards the networks when their predictions show a coarse similarity with the label of the dataset. The α and β parameters were set to (0.2, 0.8) initially, in order to counteract the above-mentioned class imbalance by weighting less the false negatives (FN, related to voxels erroneously considered to have "no material") and more the false positives (FP) in the loss function, when compared to the Sørensen–Dice loss ($\alpha = \beta = 0.5$). For a more in-depth comparison, the networks were also trained with the α and β parameters equal to (0.8, 0.2), so that it could be possible to investigate how they affect the final accuracy of the networks.

The networks architecture and the training/validation/testing procedure were developed using PyTorch libraries[22] version 1.11.0, based on CUDA[23] version 11.6, and PyTorch Lightning,[24] while the data augmentations and 3D patch-sampling patterns were implemented on the TorchIO library.[25] A unique main seed across the training governs all the extraction from a random distribution (for the random choice of the 3D patch location, which data augmentation to apply, etc.) and CUDA's deterministic algorithms were used if available. To have a constant input dimension of 64^3 for all the examples supplied to the networks, patches of identical dimension were sampled from the datasets with a uniform likelihood over all the possible patches of the dataset that have at least a positive-label in that patch, in order to reduce the class imbalance between positive and negative labels. The trainings have been executed with the same number of patches per epoch (7200) independent of the training dataset and for a total of 150 epochs and a constant learning rate of 0.0001. In order to ensure a batch size of 360 across all the networks while ensuring the different memory requirements of the different networks are met, the gradients of multiple batches were accumulated before backpropagation.

3. RESULTS & DISCUSSION

The validation Sørensen–Dice scores for all the trainings are shown in Figs. 4a and 4b, as the mean value among all the models that share the same training dataset (Fig. 4a) or the same typology of network (Fig. 4b). From



(a) (b)

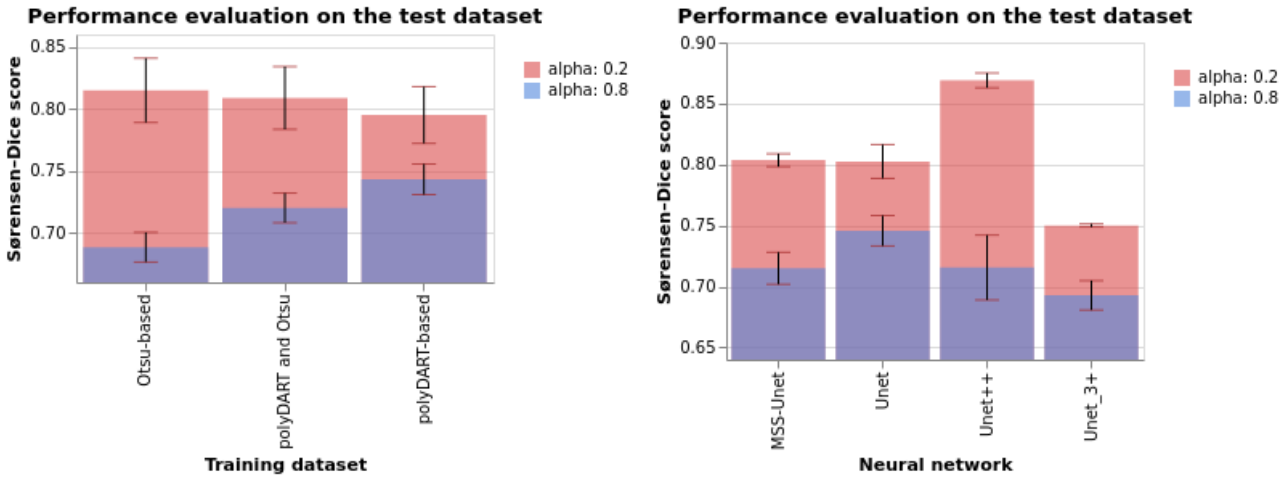
Figure 4: Bar chart with the average score achieved by the networks on the validation dataset at the end of the training. The error-bars represent the standard deviation of the mean. Results are shown for the two alpha values of the Focal Tversky function and grouped by the training sets (left) and network typology (right).

the validation perspective, an improvement in performance is noticeable when switching from a loss function that is more sensitive to defects (alpha 0.2, beta 0.8) to a less defect-sensible loss function (alpha 0.8, beta 0.2). The best network on the validation dataset is the UNet 3+ (when trained with the "polyDART and Otsu" dataset), for which a visual example of inference is shown in Fig. 3. Among the networks, the score is not significantly affected by the architecture of the networks - with the exception of the UNet++, which shows a considerably worse performance.

Results for the application of the network on the noisy cubic sample that has been manually segmented are shown in Figure 5b and 5a. Interestingly, UNet++ shows the best performance on this test dataset (Fig. 5b), as it is more capable of mimicking the human segmentation pattern than the other networks. In Figure 6 it is shown a region of interest of 50x50 pixels within the noisy sample, alongside its manual segmentation, the output of UNet++ and the less accurate output of the best network of the validation dataset (UNet 3+). As evident from this figure or Figure 5b and 5a, better performances are achieved with a lower value of the α parameter. This apparent clash between the performances of the networks on the validation and the test datasets is explainable by considering the lower contrast of pores of the test dataset and that during the training/validation the Sørensen-Dice score is calculated against the labels produced by two segmentation techniques that were not sensitive to pores as the human operator would; higher Sørensen-Dice score are in fact achieved when the networks are trained to be less sensitive in this case. Regarding the performance of the networks on the test dataset due to the particular training datasets, the high standard deviation of the mean (Fig. 5a) does not allow concluding that the different training-sets are impacting the performance in a significant way.

4. CONCLUSIONS

Several recent deeply supervised networks have been trained and evaluated for the pore segmentation task on X-CT images of AM samples, with a 3D patch-based approach. The segmentation task has been shown to be heavily dependent on the parameters of the FTL function that had been used for training. In particular, higher values of α (and lower of β) make the networks more sensible to pores, which is particularly appreciated during the application of these networks on the testing dataset, where pores have smaller contrast due to severe artefacts. Among all the networks, UNet++ (with average Sørensen–Dice score of 0.8688 $\pm$ 0.006) had been capable of achieving a better performance on the dataset characterised by severe artefacts, by being more consistent with the segmentation performed by a human operator, when compared with other networks (UNet-MSS: 0.8035 $\pm$ 0.005, UNet: 0.8024 $\pm$ 0.013 and UNet 3+: 0.75 $\pm$ 0.0013). When the Sørensen–Dice score is evaluated on the validation dataset, for which the labels were obtained with the same techniques used for the training dataset,



(a) (b)

Figure 5: Bar chart with the average score achieved by the networks on the test dataset at the end of the training. The error-bars represent the standard deviation of the mean. Results are shown for the two alpha values of the Focal Tversky function and grouped by the training sets (left) and network typology (right).

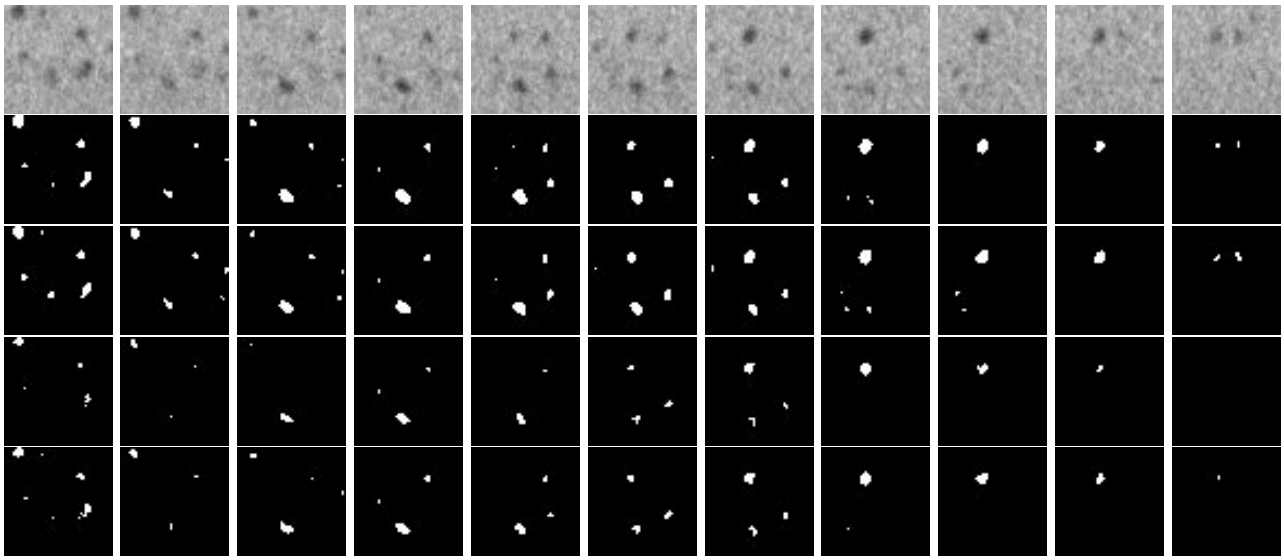


Figure 6: Comparison of a 50x50 pixel patch (500x500 µm, as highlighted in Fig. 2) extracted from the testing dataset for 11 consecutive slices: On the first row it is shown the FDK reconstruction, on the second it is shown the manual segmentation, on the third row the prediction of the UNet++, on the forth row the prediction of the UNet 3+ (with the parameters that lead to the best performance on the validation dataset) and on the fifth row the prediction of the UNet 3+ (with the parameters that lead UNet++ to perform better on the test dataset).

then the results indicate that UNet 3+ (0.9248 ± 0.0035) and UNet-MSS (0.9227 ± 0.0026) perform best, as they were able to better capture how the two segmentation methods work. Lastly, the analysis on the dependence of the networks' performance on the training datasets, did not show a correlation with the achievement of a better performance. This result may indicate that the information taught to the networks by the different training-sets is redundant or that the training-sets are not big enough to teach the networks misbehaviours of one or another of the two segmentation techniques. A natural future development of this work is expand the comparison by inspecting the performance of these networks for other meaningful values of the FTL parameters' α and β; the optimal values are linked to the contrast of pores, but a clear connection is yet to be established. Another

research path that could follow this work, is the comparison between the performance of unsupervised networks and the deeply supervised on a limited dataset for the segmentation task. Unsupervised learning techniques are appealing because they do not necessarily require a dataset of real samples, but at the same time they heavily depend on the fidelity of the synthetic training-set - along with noise and X-CT artefacts - to a real training-set. Meanwhile, the training of supervised networks requires annotated data that may not be available. With these technical difficulties in mind, the choice of one or the other method could then depend on the desired performance.

5. ACKNOWLEDGEMENTS

This work is financially supported by the VLAIO ICON project VIL (HBC.2019.2808) and the Research Foundation - Flanders (FWO) (S004217N, S003421N) and the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" programme.

REFERENCES

[1] Zhang, B., Li, Y., and Bai, Q., "Defect formation mechanisms in selective laser melting: a review," *Chinese Journal of Mechanical Engineering* **30**(3), 515–527 (2017).

[2] Leary, M., Mazur, M., Elambasseril, J., McMillan, M., Chirent, T., Sun, Y., Qian, M., Easton, M., and Brandt, M., "Selective laser melting (SLM) of AlSi12Mg lattice structures," *Materials & Design* **98**, 344–357 (2016).

[3] Thompson, A., Maskery, I., and Leach, R. K., "X-ray computed tomography for additive manufacturing: a review," *Measurement Science and Technology* **27**(7), 072001 (2016).

[4] Sanaei, N. and Fatemi, A., "Defects in additive manufactured metals and their effect on fatigue performance: a state-of-the-art review," *Progress in Materials Science* **117**, 100724 (2021).

[5] Heylen, R., Thanki, A., Verhees, D., Iuso, D., De Beenhouwer, J., Sijbers, J., Witvrouw, A., Haitjema, H., and Bey-Temsamani, A., "3D total variation denoising in X-CT imaging applied to pore extraction in additively manufactured parts," *Measurement Science and Technology* **33**(4), 045602 (2022).

[6] Du Plessis, A., Sperling, P., Beerlink, A., Tshabalala, L., Hoosain, S., Mathe, N., and Le Roux, S. G., "Standard method for microCT-based additive manufacturing quality control 1: Porosity analysis," *MethodsX* **5**, 1102–1110 (2018).

[7] Gobert, C., Kudzal, A., Sietins, J., Mock, C., Sun, J., and McWilliams, B., "Porosity segmentation in X-ray computed tomography scans of metal additively manufactured specimens with machine learning," *Additive Manufacturing* **36**, 101460 (2020).

[8] Wong, V. W. H., Ferguson, M., Law, K. H., Lee, Y.-T. T., and Witherell, P., "Automatic volumetric segmentation of additive manufacturing defects with 3D U-Net," *arXiv preprint arXiv:2101.08993* (2021).

[9] Six, N., De Beenhouwer, J., and Sijbers, J., "Poly-DART: A discrete algebraic reconstruction technique for polychromatic x-ray CT," *Optics Express* **27**(23), 33670–33682 (2019).

[10] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[11] Zhao, W., Jiang, D., Queralta, J. P., and Westerlund, T., "MSS U-Net: 3D segmentation of kidneys and tumors from CT images with a multi-scale supervised U-Net," *Informatics in Medicine Unlocked* **19**, 100357 (2020).

[12] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J., "UNet++: A nested u-net architecture for medical image segmentation," in [*Deep learning in medical image analysis and multimodal learning for clinical decision support*], 3–11, Springer (2018).

[13] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J., "Unet 3+: A full-scale connected unet for medical image segmentation," in [*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 1055–1059, IEEE (2020).

[14] De Samber, B., Renders, J., Elberfeld, T., Maris, Y., Sanctorum, J., Six, N., Liang, Z., De Beenhouwer, J., and Sijbers, J., "FleXCT: a flexible X-ray CT scanner with 10 degrees of freedom," *Optics Express* **29**(3), 3438–3457 (2021).

[15] Feldkamp, L. A., Davis, L. C., and Kress, J. W., "Practical cone-beam algorithm," *Josa a* **1**(6), 612–619 (1984).

[16] Van Aarle, W., Palenstijn, W. J., Cant, J., Janssens, E., Bleichrodt, F., Dabravolski, A., De Beenhouwer, J., Batenburg, K. J., and Sijbers, J., "Fast and flexible X-ray tomography using the ASTRA toolbox," *Optics express* **24**(22), 25129–25147 (2016).

[17] Iuso, D., Nazemi, E., Six, N., De Samber, B., De Beenhouwer, J., and Sijbers, J., "CAD-Based scatter compensation for polychromatic reconstruction of additive manufactured parts," in [*2021 IEEE International Conference on Image Processing (ICIP)*], 2948–2952, IEEE (2021).

[18] Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G., "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[19] He, K., Zhang, X., Ren, S., and Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in [*Proceedings of the IEEE international conference on computer vision*], 1026–1034 (2015).

[20] Glorot, X. and Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," in [*Proceedings of the thirteenth international conference on artificial intelligence and statistics*], 249–256, JMLR Workshop and Conference Proceedings (2010).

[21] Abraham, N. and Khan, N. M., "A novel focal tversky loss function with improved attention U-Net for lesion segmentation," in [*2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*], 683–687, IEEE (2019).

[22] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems* **32** (2019).

[23] NVIDIA, P. V. and Fitzek, F., "Cuda, release: 10.2. 89, 2020," (2020).

[24] Falcon, W., Borovec, J., Wälchli, A., Eggert, N., Schock, J., Jordan, J., Skafte, N., Bereznyuk, V., Harris, E., Murrell, T., et al., "Pytorchlightning/pytorch-lightning: 0.7. 6 release," *Zenodo: Geneva, Switzerland* (2020).

[25] Pérez-García, F., Sparks, R., and Ourselin, S., "TorchIO: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine* **208**, 106236 (2021).