EPiC Series in Health Sciences

Volume 5, 2022, Pages 31–36

Proceedings of The 20th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

Transformer vs. CNN – A Comparison on Knee Segmentation in Ultrasound Images

Peter Brößner[1], Benjamin Hohlmann[1], Klaus Radermacher[1]

[1] Chair of Medical Engineering, Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Aachen, 52074, Germany.

broessner@hia.rwth-aachen.de

Abstract

The automated and robust segmentation of bone surfaces in ultrasound (US) images can open up new fields of application for US imaging in computer-assisted orthopedic surgery, e.g. for the patient-specific planning process in computer-assisted knee replacement. For the automated, deep learning-based segmentation of medical images, CNN-based methods have been the state of the art over the last years, while recently Transformer-based methods are on the rise in computer vision. To compare these methods with respect to US image segmentation, in this paper the recent Transformer-based Swin-UNet is exemplarily benchmarked against the commonly used CNN-based nnUNet on the application of in-vivo 2D US knee segmentation.

Trained and tested on our own dataset with 8166 annotated images (split in 7155 and 1011 images respectively), both the nnUNet and the pre-trained Swin-UNet show a Dice coefficient of 0.78 during testing. For distances between skeletonized labels and predictions, a symmetric Hausdorff distance of 44.69 pixels and a symmetric surface distance of 5.77 pixels is found for nnUNet as compared to 42.78 pixels and 5.68 pixels respectively for the Swin-UNet. Based on qualitative assessment, the Transformer-based Swin-UNet appears to benefit from its capability of learning global relationships as compared to the CNN-based nnUNet, while the latter shows more consistent and smooth predictions on a local level, presumably due to the character of convolution operation. Besides, the Swin-UNet requires generalized pre-training to be competitive.

Since both architectures are evenly suited for the task at hand, for our future work, hybrid architectures combining the characteristic advantages of Transformer-based and CNN-based methods seem promising for US image segmentation.

1 Introduction

Ultrasound (US) imaging is a cost-efficient, mobile and non-invasive imaging modality, which has the drawback of challenging interpretability due to imaging artifacts and noise. For the widespread use

F. Rodriguez Y Baena, J.W. Giles and E. Stindel (eds.), CAOS 2022 (EPiC Series in Health Sciences, vol. 5), pp. 31–36

of US in applications of computer-assisted orthopedic surgery, the robust segmentation of bone surfaces can address this drawback and thus open up new areas of application. It could establish US as an alternative to computed tomography (CT) imaging, e.g. for the patient-specific planning process in computer-assisted knee replacement; here, the use of CT imaging comes with high cost, lengthy waiting times and exposes the patient to harmful radiation (Bae and Song 2011). Automating the process of segmentation offers the advantages of being more time efficient and consistent as compared to manual expert segmentation (Kubicek et al. 2019). This automation of segmentation can be realized using deep learning-based approaches, which incorporate expert knowledge in the learning process; the feasibility of automating US segmentation via deep learning has been shown in previous publications (Hohlmann et al. 2021).

For deep learning-based segmentation, approaches based on convolutional neural networks (CNNs) defined the state of the art over the last years (Minaee et al. 2021). But most recently, Transformer-based architectures are on the rise in computer vision, inspired by their outstanding performance in the field of language processing (Khan et al. 2022). While CNNs are powerful in learning local patterns, they are less suited for learning global, image-wide relationships due to the character of convolution operation (Cao et al. 2021). Transformers on the other hand are based on the self-attention mechanism, which is able to capture context over the complete input, thus being capable of learning global relationships. Moreover, since Transformers assume minimum prior knowledge of a problem, they benefit from pre-training on non-specialized, large-scale datasets (Khan et al. 2022).

The contribution of this paper is an exemplary comparison of a recent Transformer-based architecture with an established and commonly used CNN-based method on the application of in-vivo 2D knee bone segmentation in US images.

2 Materials and Methods

For the task of medical image segmentation, encoder-decoder architectures with skip connections as the UNet (Ronneberger et al. 2015) are most commonly used. Thus, two architectures with similar UNet-like structures are analyzed. Furthermore, we choose a pure Transformer-based and a pure CNN-based architecture in order to allow a distinct comparison, excluding hybrid architectures. As Transformer-based architecture, the Swin-UNet (Cao et al. 2021) is selected, since it employs a pure state of the art Swin Transformer backbone (Liu et al. 2021) and achieves state of the art performance in the Automatic Cardiac Diagnosis Challenge (ACDC) (Bernard et al. 2018). As CNN-based architecture, the nnUNet (Isensee et al. 2021) is used, as it offers a self-configuring framework and is intended to be utilized as a reference. Since the Swin-UNet is only available for 2D segmentation, the 2D variant of the nnUNet is employed.

For the training of both architectures, a dataset consisting of 8166 annotated in-vivo ultrasound images of the knee is used, split in subsets of 7155 images for training and 1011 images for testing; each image has an equal size of 384x384 pixels. The dataset contains images of different young, mostly male test subjects, acquired on two devices (*Clarius L15 HD* and *SonixTouch Q+* equipped with 2D and 3D probe) by different operators.

For the training of the Swin-UNet, a rudimentary grid search for the hyper-parameters of learning rate, loss weighting and window size is carried out for optimization. As proposed by (Cao et al. 2021), we resize the input of Swin-UNet to 224x224 pixels for improved efficiency. Furthermore, we train one variant with initial weights pre-trained on ImageNet-1K, as provided by (Liu et al. 2021). For the nnUNet, the optimization of hyper-parameters is automated by the framework. All models are trained using the maximum available batch size for about 24 hours, resulting in 150 (Swin-UNet) and 325 (nnUNet) epochs; the latest models are used for evaluation respectively.

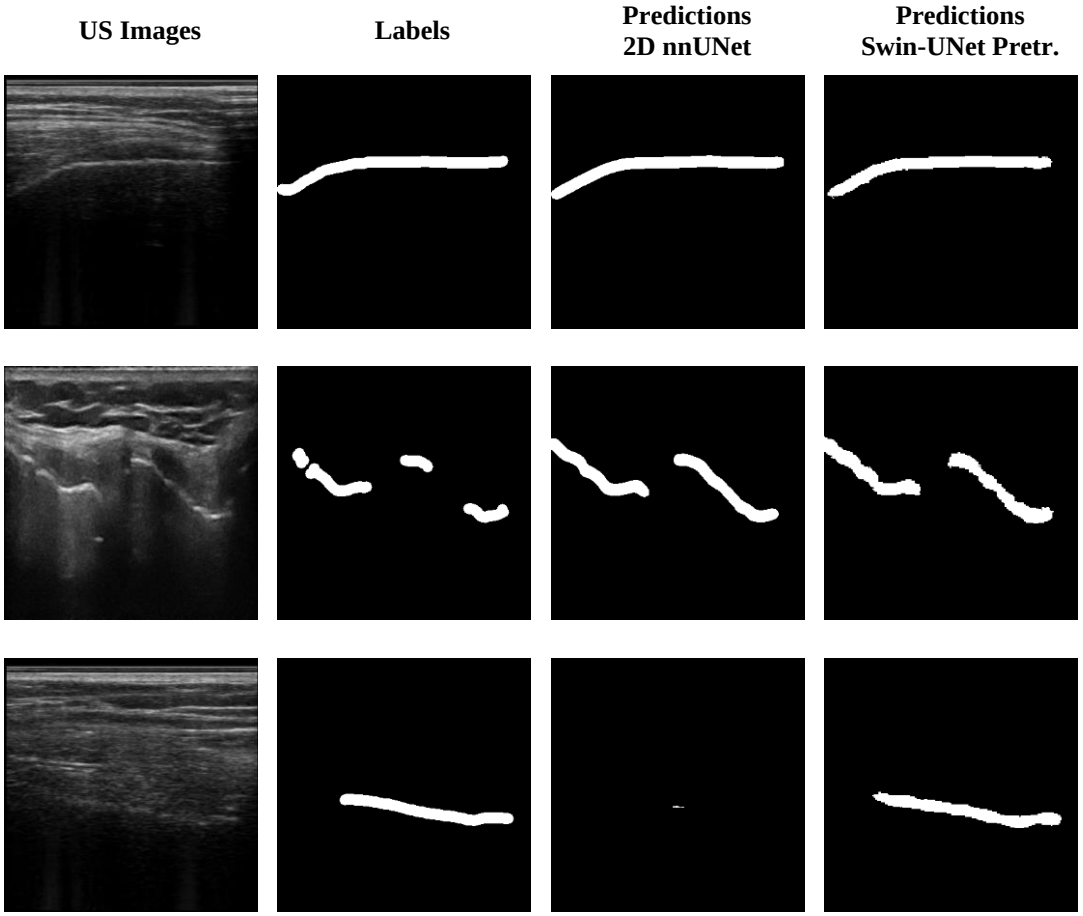
3 Results

For a comparison of segmentation results, Dice coefficient (DC) is used as a pixel-based metric; additionally, distances between the center lines of labels and predictions are measured by means of symmetrical surface distance (SSD) and symmetrical Hausdorff distance (SHD). The comparison of results can be seen in Table 1.

Architecture	Dice Coefficient	Sym. Surface Distance	Sym. Hausdorff Distance
2D nnUNet	0.78	5.77 pixels	44.69 pixels
Swin-UNet	0.60	29.95 pixels	147.35 pixels
Swin-UNet Pre-trained	0.78	5.68 pixels	42.78 pixels

Table 1: Comparison of segmentation results for 2D nnUNet and Swin-UNet.

Furthermore, an exemplary qualitative comparison of segmentation results is shown in Figure 1, with US images in the first column, corresponding annotations in the second column and predictions of 2D nnUNet (third column) and pre-trained Swin-UNet (fourth column).



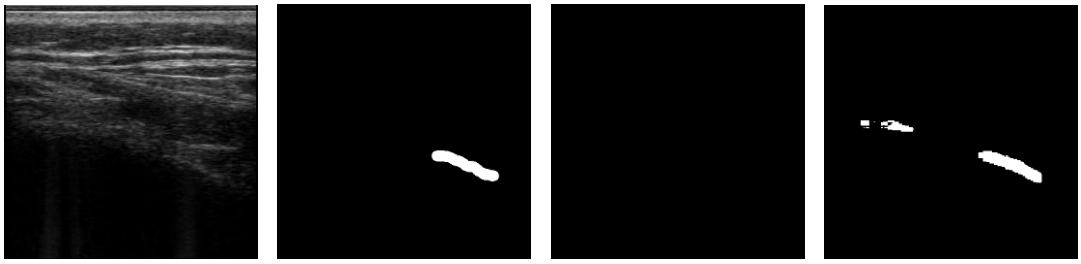


Figure 1: Qualitative comparison of US images (1st column), labels (2nd column), segmentation results of 2D nnUNet (3rd column) and pre-trained Swin-UNet (4th column).

4 Discussion

A comparison of segmentation results between the Transformer-based Swin-UNet models first shows the significant impact of pre-training, with an improvement of DC by about 30% and a reduction of SHD by about 81%. Segmentation results of CNN-based 2D nnUNet and pre-trained Swin-UNet, on the other hand, are only slightly different: While both achieve about equal results by means of DC and SSD, the pre-trained Swin-UNet shows a small advantage in terms of SHD. This indicates Swin-UNet segmentations covering a bigger part of the bone, an assumption that is substantiated by the qualitative results: The 2D nnUNet shows a tendency to rather conservative segmentations while the Swin-UNet shows a tendency to more (over-) complete segmentations, maybe due to learning global relationships (see rows 3 and 4 of Figure 1). Segmentations of the 2D nnUNet seem more consistent and smooth on a local level as opposed to rather frayed segmentations of the Swin-UNet. Besides, both models show their general capability of segmenting bone in US images (see row 1 of Figure 1), sometimes apparently even outperforming the human annotator (see row 2 of Figure 1); the rather conservative annotation evident in this example may also lead to an underestimation in terms of segmentation metrics.

Images contained in our dataset have a maximum isotropic pixel spacing of 0.1mm, so for the pre-trained Swin-UNet, we find a SSD of 0.57mm, which is about the in-slice resolution of CT, and a SHD of 4.28mm. In comparison to related work on US segmentation, (Hohlmann et al. 2020) achieved very similar results for US femur segmentation, with a reported SSD of 0.63mm and a SHD of 4.2mm. For a different application, namely the segmentation of alveolar bone in intra-oral US images, (Duong et al. 2019) found a Dice coefficient of 0.75, as compared to 0.78 for Swin-UNet and nnUNet in our case.

Our study is limited by the non-representative composition of the dataset we used for training and evaluation. The chosen split into subsets for training and testing may also have affected results. Furthermore, we made no changes to the original, optimized architectures, which means that structures of the employed networks are not identical.

5 Conclusion

Both, CNN-based 2D nnUNet and Transformer-based pre-trained Swin-UNet show their capability of segmenting knee bones in US images, with the latter significantly benefitting from generally pre-trained weights. Qualitatively, the Swin-UNet shows more complete predictions of the bone surface, likely due to its global interpretation capabilities, while the nnUNet shows more consistent and smooth predictions on a local level. Overall, both architectures seem evenly suited for the task at hand, both showing characteristic advantages for their respective method; thus, hybrid architectures combining the

advantages of Transformer-based and CNN-based methods may be most promising for US segmentation.

6 Acknowledgements

Simulations were partly performed with computing resources granted by RWTH Aachen University, under project rwth0536.

References

Bae, Dae Kyung; Song, Sang Jun (2011): Computer assisted navigation in knee arthroplasty. In *Clin Orthop Surg* 3 (4), pp. 259–267. DOI: 10.4055/cios.2011.3.4.259.

Bernard, Olivier; Lalande, Alain; Zotti, Clement; Cervenansky, Frederick; Yang, Xin; Heng, Pheng-Ann et al. (2018): Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? In *IEEE Trans. Med. Imaging* 37 (11), pp. 2514–2525. DOI: 10.1109/tmi.2018.2837502.

Cao, Hu; Wang, Yueyue; Chen, Joy; Jiang, Dongsheng; Zhang, Xiaopeng; Tian, Qi; Wang, Manning (2021): Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation.

Duong, Dat Q.; Nguyen, Kim-Cuong T.; Kaipatur, Neelambar R.; Lou, Edmond H. M.; Noga, Michelle; Major, Paul W. et al. (2019): Fully Automated Segmentation of Alveolar Bone Using Deep Convolutional Neural Networks from Intraoral Ultrasound Images. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE.

Hohlmann, Benjamin; Brößner, Peter; Welle, Kristian; Radermacher, Klaus (2021): Segmentation of the Scaphoid Bone in Ultrasound Images. In *Current Directions in Biomedical Engineering* (1), pp. 76–80.

Hohlmann, Benjamin; Glanz, Jakob; Radermacher, Klaus (2020): Segmentation of the distal femur in ultrasound images. In *Current Directions in Biomedical Engineering* 6 (1).

Isensee, Fabian; Jaeger, Paul F.; Kohl, Simon A. A.; Petersen, Jens; Maier-Hein, Klaus H. (2021): nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. In *Nat Methods* 18 (2), pp. 203–211. DOI: 10.1038/s41592-020-01008-z.

Khan, Salman; Naseer, Muzammal; Hayat, Munawar; Zamir, Syed Waqas; Khan, Fahad Shahbaz; Shah, Mubarak (2022): Transformers in Vision: A Survey. In *ACM Comput. Surv.* DOI: 10.1145/3505244.

Kubicek, Jan; Tomanec, Filip; Cerny, Martin; Vilimek, Dominik; Kalova, Martina; Oczka, David (2019): Recent Trends, Technical Concepts and Components of Computer-Assisted Orthopedic Surgery Systems: A Comprehensive Review. In *Sensors (Basel, Switzerland)* 19 (23), p. 5199. DOI: 10.3390/s19235199.

Liu, Ze; Lin, Yutong; Cao, Yue; Hu, Han; Wei, Yixuan; Zhang, Zheng et al. (2021): Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV): IEEE.

Minaee, Shervin; Boykov, Yuri Y.; Porikli, Fatih; Plaza, Antonio J.; Kehtarnavaz, Nasser; Terzopoulos, Demetri (2021): Image Segmentation Using Deep Learning: A Survey. In *IEEE Trans. Pattern Anal. Mach. Intell.* PP, p. 1. DOI: 10.1109/TPAMI.2021.3059968.

Ronneberger, Olaf; Fischer, Philipp; Brox, Thomas (2015): U-Net: Convolutional Networks for Biomedical Image Segmentation. In, vol. 9351. International Conference on Medical Image Computing and Computer-Assisted Intervention: Springer, Cham, pp. 234–241.