EPiC Series in Health Sciences

Volume 8, 2026, Pages 152–155

Proceedings of The 25th Annual Meeting of the International Society for Computer Assisted Orthopaedic Surgery

From Unlabeled Data to Clinical Applications: Foundation Models in Medical Imaging

Joshua Scheuplein[1,2], Maximilian Rohleder[1,2], Björn Kreher[2], and Andreas Maier[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Advanced Therapies, Siemens Healthineers AG, Forchheim, Germany
joshua.scheuplein@fau.de

Abstract

The performance of deep learning algorithms is highly dependent on the quantity and diversity of the available training data. However, obtaining sufficiently large datasets represents a significant challenge, particularly in the field of medical imaging. This study underscores the potential of self-supervised training strategies in the development of deep learning models for medical imaging tasks. It is demonstrated that workflows can be significantly optimized by incorporating the feature content of a large collection of medical X-ray images from intraoperative C-arm scans into a so-called foundation model. This approach facilitates the efficient adaptation to a variety of concrete applications by fine-tuning a small task-specific head network on top of the pre-trained foundation model, thereby reducing both computational demands and training time.

1 Introduction

In recent years, the application of deep learning to medical imaging has achieved remarkable success, often surpassing the performance of traditional methods [1]. However, for these models to transition from research to clinical practice, they must meet rigorous standards of reliability, generalizability, and diagnostic accuracy [1]. A major obstacle to achieving robust performance in clinical settings is the limited availability of sufficiently large training datasets, as the size and diversity of the data are directly related to the effectiveness of a trained model [2].

Foundation models offer a promising solution to this limitation by encoding general domain knowledge that can be easily adapted to a variety of downstream tasks with minimal additional fine-tuning [3]. In particular, self-supervised learning techniques have emerged as powerful tools in this context, providing capabilities to take advantage of the vast amount of unlabeled clinical data available in many hospitals [4]. Given this background, it is demonstrated how self-supervised pre-training of feature extraction backbones on a large set of unlabeled medical X-ray images from intraoperative C-arm scans can effectively support the concurrent solution of classification, segmentation, and detection problems.

2 Materials and Methods

The general workflow followed in this study is illustrated in Figure 1. Initially, a comprehensive dataset consisting of 632,385 X-ray images was curated from intraoperative C-arm scans provided by multiple clinical collaboration partners. The images were recorded using various acquisition devices and imaging protocols, including 3D projection and fluoroscopic imaging for example. The raw data was converted to a uniform format and anonymized before being stored in a secure data lake, ensuring that no identifiable patient information could be derived.

Given the absence of additional label annotations for the images, the self-supervised learning framework DINO (i.e., knowledge-**di**stillation with **no** labels), developed by Caron et al., was employed [4]. The DINO method is based on a student-teacher network architecture and was used to train different backbone architectures for feature extraction. Subsequently, the trained backbone models were evaluated on specific downstream tasks to assess the quality of the learned feature representations. Specifically, the performance of the models was tested on body region classification, metal implant segmentation, and screw object detection. For each task, only a small, task-specific head network was trained, while the backbone models remained frozen. This approach exploits the domain knowledge already learned during the backbone pre-training, thus requiring significantly less time and computational resources to adapt to the specific task.

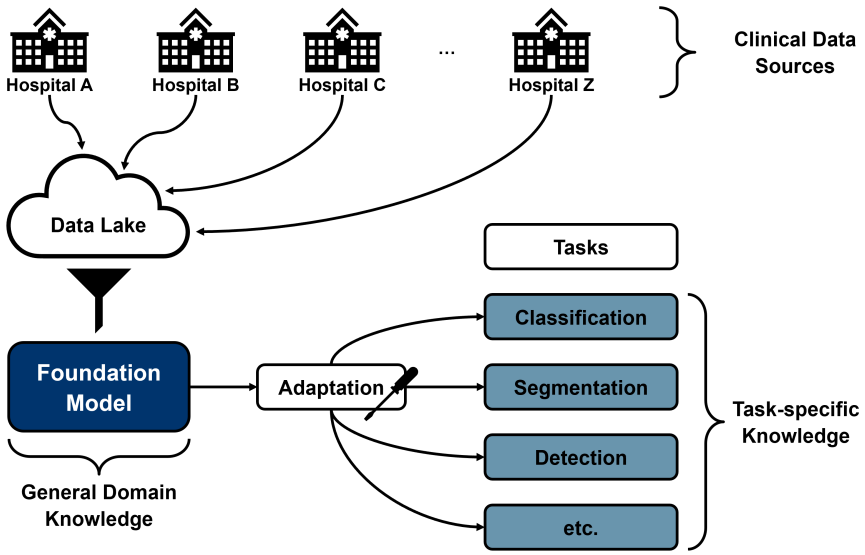

Figure 1: Workflow overview for pre-training and fine-tuning a foundation model.

3 Results

Figure 2 illustrates the inference results for the previously mentioned downstream tasks on a representative test sample showing a clinical spine scan with additional screw implants. On the left, the three body region labels with the highest probability scores out of 11 possible categories are shown. Additionally, the predicted segmentation mask highlighting metallic areas in the image as well as the detected screw bounding boxes together with their respective confidence scores are visualized.

To complement this qualitative demonstration, a quantitative evaluation was also performed to confirm the reliable adaptation of the backbone models. Body region classification was best solved using a vision transformer (ViT) backbone with an accuracy of 96.9%, while precision and recall reached 97.0% each. In contrast, backbones based on residual network (ResNet) architectures delivered the best performance in the segmentation task with an average DICE score of 94.1%. Moreover, screw objects were detected with an average deviation of 3.28 and 7.86 pixels for the head and tip coordinates, respectively, while the average angular error between ground truth and prediction was 1.22 degrees.

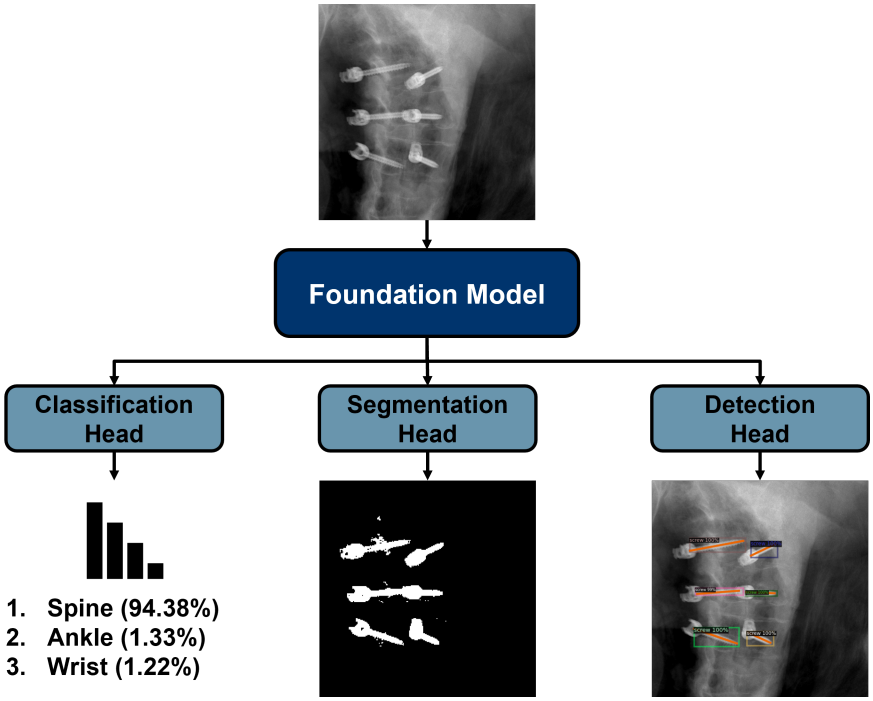

Figure 2: Exemplary inference results for a classification, segmentation, and detection task.

4 Discussion and Conclusion

This study demonstrates the efficacy of self-supervised learning for pre-training foundation models in medical imaging, aligning with prior research emphasizing the importance of self-supervised strategies for domains with limited annotated data [5, 6]. Notably, the results reveal substantial differences in the performance of the ViT and ResNet backbones. Specifically, the ViT model demonstrated superior classification accuracy, while the ResNet-based backbone exhibited better performance in the segmentation and detection tasks. These outcomes can be attributed to the architectural strengths of each model. ViTs are advantageous in capturing global contextual information due to their attention-based mechanisms, making them particularly effective for classification tasks that require holistic image understanding [7, 8]. Conversely, ResNet architectures, with their hierarchical feature extraction and strong spatial localization capabilities, are better suited for tasks requiring precise pixel-level predictions [9, 10].

Despite these achievements, several challenges persist. The evaluation of the pre-trained feature extraction models was conducted on only three downstream tasks, which may not fully cover the diverse range of clinical application scenarios. Additionally, while the frozen backbone approach offers computational efficiency, it may limit the adaptability of the model for tasks that require a deeper contextual understanding. In conclusion, this study contributes to the advancement of foundation models in medical imaging, providing a scalable solution to the challenge of missing label annotations. The potential of this approach lies in its ability to streamline the development of novel, task-specific models, thereby accelerating the integration of deep learning solutions into clinical workflows.

References

[1] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. Deep Learning in Medical Image Analysis. In *Deep Learning in Medical Image Analysis*, volume 1213, pages 3–21. Springer International Publishing, Cham, 2020. Series Title: Advances in Experimental Medicine and Biology.

[2] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[3] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision, October 2023. arXiv:2310.18689 [cs].

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, Montreal, QC, Canada, October 2021. IEEE.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, and others. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[7] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do Vision Transformers See Like Convolutional Neural Networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12116–12128. Curran Associates, Inc., 2021.

[8] Zhendong Liu, Shuwei Qian, Changhong Xia, and Chongjun Wang. Are transformer-based models more robust than CNN-based models? *Neural Networks*, 172:106091, April 2024.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.