DINO Adapted to X-Ray (DAX): Foundation Models for Intraoperative X-Ray Imaging

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

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

Abstract. Intraoperative X-ray imaging represents a key technology for guiding orthopedic interventions. Recent advancements in deep learning have enabled automated image analysis in this field, thereby streamlining clinical workflows and enhancing patient outcomes. However, many existing approaches depend on task-specific models and are constrained by the limited availability of annotated data. In contrast, self-supervised foundation models have exhibited remarkable potential to learn robust feature representations without label annotations. In this paper, we introduce **D**INO **A**dapted to **X**-ray **(DAX)**, a novel framework that adapts DINO for training foundational feature extraction backbones tailored to intraoperative X-ray imaging. Our approach involves pre-training on a novel dataset comprising over 632,000 image samples, which surpasses other publicly available datasets in both size and feature diversity. To validate the successful incorporation of relevant domain knowledge into our DAX models, we conduct an extensive evaluation of all backbones on three distinct downstream tasks and demonstrate that small head networks can be trained on top of our frozen foundation models to successfully solve applications regarding (1) body region classification, (2) metal implant segmentation, and (3) screw object detection. The results of our study underscore the potential of the DAX framework to facilitate the development of robust, scalable, and clinically impactful deep learning solutions for intraoperative X-ray image analysis. Source code and model checkpoints are available at https://github.com/JoshuaScheuplein/DAX.

Keywords: Deep learning · Self-supervised learning · Foundation model · Intraoperative X-ray imaging

1 Introduction

Medical imaging plays a crucial role in modern healthcare, enabling the diagnosis and treatment of various conditions [24]. In the field of orthopedic interventions, intraoperative X-ray imaging stands out because of its widespread accessibility, cost-effectiveness, and the ability to provide real-time visualization of anatomical structures [1]. However, the automated analysis of intraoperative X-ray images remains challenging due to the diverse range of imaging protocols, variability

in patient anatomy, and the presence of artifacts [27]. In recent years, deep learning has emerged as a promising technique for developing advanced image analysis algorithms that address these limitations [5]. Nevertheless, several challenges persist, including the limited availability of training data, variations in input, problems with context understanding, and the fact that most models are designed for specific tasks only [21].

Given this context, foundation models have seen increased usage in this area to acquire general domain knowledge that can be subsequently reused for a variety of downstream tasks with minimal additional fine-tuning [2]. For instance, Kirillov et al. introduced the Segment Anything Model (SAM) for generating object segmentation masks in a given image based on various segmentation prompts [12]. While the original SAM model primarily targets real-world images, it has been frequently adapted for medical images, such as MedSAM [16]. In general, existing literature indicates a shift towards adapting natural domain foundation models explicitly for medical imaging applications, given their often limited transferability [3,9]. In the field of medical X-ray imaging, for instance, Shakouri et al. developed DINO-CXR [22], which is primarily trained on chest X-ray (CXR) data. However, chest and other radiological X-ray images do not capture the same feature content as intraoperative scans that are commonly used in orthopedic procedures. The latter exhibit greater variability across body regions and less strict standard views than diagnostic imaging, while introducing additional complexity due to diverse implant types and imaging characteristics, such as increased scatter radiation [27].

Therefore, we propose **D**INO **A**dapted to **X**-ray **(DAX)** that translates the capabilities of the so-called knowledge **di**stillation with **no** labels **(DINO)** framework to the domain of intraoperative X-ray imaging. DINO [4] and its revised version, DINOv2 [18], represent self-supervised training methods that are particularly well-suited for the domain of medical imaging, where label annotations are typically scarce. Our pre-training dataset exceeds the scale and feature diversity of publicly available datasets, such as MIMIC-CXR [11] and CheXpert [10]. We pre-train several backbone architectures, including residual networks (ResNets) [8] and vision transformers (ViTs) [6] on this dataset with DAX. Our findings demonstrate the capacity of these models to incorporate meaningful domain knowledge, which can be used for solving clinically relevant downstream tasks. We summarize our contributions as follows:

1. **Dataset Collection.** A comprehensive and diverse dataset comprising over 632,000 intraoperative X-ray images collected from different sources is used for model pre-training.
2. **Methodological Development.** We adapt DINO for the field of intraoperative X-ray imaging, incorporating domain-specific image preprocessing and data augmentation strategies.
3. **Model Training and Evaluation.** Various backbone architectures are pre-trained using the novel DAX framework and extensively evaluated on multiple downstream tasks including body region classification, metal implant segmentation, and screw object detection.

2 Materials and Methods

Dataset. The pre-training dataset for the DAX foundation models comprises 632,385 images in total. These originate from scans of human cadaveric specimens as well as trauma and orthopedic surgeries, acquired using both fixed and mobile C-arm systems. Table 1 summarizes the relative proportion of anatomical regions represented in the dataset. The category "upper extremities" includes wrist, elbow, and shoulder scans; "lower extremities" cover foot, leg, and pelvis.

Table 1. Body region distribution in the pre-training dataset.

	# Images	Broncho	Spine	Upper Extremities	Lower Extremities
Cadaver	514,577	0.0%	16.5%	39.2%	44.2%
Clinical	117,808	5.9%	33.7%	14.7%	45.7%
Total	632,385	1.1%	19.7%	34.7%	44.5%

DAX Pipeline. As illustrated in Figure 1, the development pipeline for our foundation models adopts the student–teacher architecture from the original DINO implementation, while introducing two distinct approaches to image preprocessing and data augmentation specifically designed for intraoperative X-ray imaging. In version A, intensity normalization, a negative logarithm transform, and region-of-interest (ROI) normalization are first applied to enhance the contrast of the raw input images. This is followed by several image augmentation techniques, including random cropping, flipping, color jittering, and Gaussian blurring, to generate two global and multiple local crop images. In version B, the intensity normalization and the negative logarithm transform are applied at random with a probability of 0.5 to further increase data variability during pre-training. Additionally, rotation is incorporated and the Gaussian filtering operation is replaced with a sharpening function that randomly applies either blurring or sharpening to the cropped images. Notably, we train all models from scratch, without initializing them from any checkpoints pre-trained on ImageNet or other public datasets.

Implementation Details. We train ResNet18, ResNet50, and both tiny as well as small ViT architectures with either 16- or 8-pixel patch sizes using DAX. All pre-training runs are conducted on 4 NVIDIA A100 GPUs, each equipped with 80GB of RAM. Every foundational feature extraction model is pre-trained on our custom dataset for 200 epochs using the AdamW optimizer, with a linear warmup applied during the first 10 epochs and subsequent cosine annealing as learning rate schedule. The global crops have a resolution of 244×244 pixels, while the local crops are resized to 96×96 pixels. All computational operations are executed with 32-bit precision and the total pre-training time ranges from 2d 22h for ResNet18 to 27d 14h for ViT-S-8.

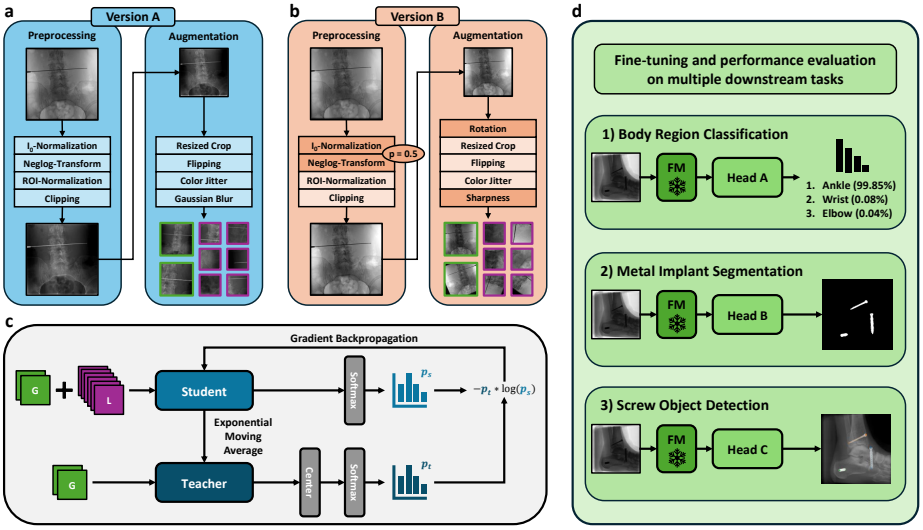

Fig. 1. Overview of the proposed DAX pipeline. (a) Initial and (b) modified image preprocessing and augmentation strategy. (c) High-level architecture of the pre-training method. (d) Overview of downstream tasks used for performance assessment.

Body Region Classification. In the first downstream experiment, we evaluate the performance of all DAX backbones in a body region classification task. The dataset consists of 2,941 annotated clinical images from intraoperative scans, which are categorized into eleven different classes and not included in the pre-training dataset. To qualitatively assess the high-dimensional feature embeddings of our foundation models, we project them into a 2D space using the uniform manifold approximation and projection (UMAP) method [17]. Additionally, we conduct quantitative analyses through linear probing on the output encodings by training a small head network comprising a single linear layer, using 5-fold cross-validation with an 80/20 train/test split, without a separate held-out dataset. The backbones remain frozen during this fine-tuning process and throughout all the other downstream tasks.

Metal Implant Segmentation. Furthermore, we evaluate the foundational feature extraction models on a metal implant segmentation task to assess their ability to capture spatial information. The feature maps from ResNet and ViT backbones are first upsampled to a resolution of 244×244 using bilinear interpolation. Subsequently, the concatenated feature maps are processed through a pixel-wise convolutional layer followed by a sigmoid activation function to generate the final mask prediction. The dataset for this task consists of 300 clinical images capturing various anatomical structures, with ground truth annotations provided by human experts. As before, we employ 5-fold cross-validation without a held-out set, using one fold for testing and the remaining folds for training.

Screw Object Detection. In the third experiment, we evaluate the DAX foundation models for object recognition and localization using a screw object detection task. For this purpose, we adapt the sparse detection transformer (Sparse DETR) framework, developed by Roh et al., and replace the default feature extractor by our frozen DAX models [20]. A simulated dataset was generated based on 57 real-world cone-beam computed tomography (CBCT) scans, covering seven different body regions, each comprising 400 single projection images. Additionally, 3D models of medical screws are projected into the 2D views using DeepDRR [25,26], with three configurations per volume containing either one, two, or three screw objects. The 171 available scenes are split in a 70/15/15 ratio for training, validation, and testing. Finally, we measure the detection performance using the object keypoint similarity (OKS), as introduced in the publicly available COCO dataset [13]. We set the screw keypoint constant κ such that an average deviation of 30 pixels - corresponding to the maximum occurring screw width in the used projection images - between ground truth and predicted keypoint locations results in an OKS of 0.5. This value represents the minimum threshold at which a prediction is considered a match.

3 Results

Body Region Classification. Figure 2 shows the 2D UMAP embeddings of all samples from the body region classification dataset obtained with a ResNet50 backbone at the beginning as well as the end of pre-training, respectively. Overall, clearly distinguishable clusters corresponding to different anatomical classes emerge. Notably, the feature encodings enable differentiation not only of the body regions but also of the perspectives, such as anterior-posterior (AP) or lateral views, at which the images are acquired.

In Table 2, we summarize the quantitative results for the linear probing experiment. The best ResNet architecture is ResNet50, which achieves an average F1-Score of 65.95%, while ViT-S-8 excels with 97.79% for the same metric among the ViT backbones. In particular, the latter ones show a good balance between precision and recall, whereas ResNets tend to prioritize retrieving all relevant samples, potentially leading to more false positives, as indicated by lower precision compared to recall values. The models trained with version B neither show significantly improved nor degraded performance in comparison to the models in version A. In summary, ViT architectures clearly outperform ResNets in this downstream task.

Metal Implant Segmentation. The results for the metal implant segmentation task are reported in Table 3. ResNet50 demonstrates superior performance compared to ResNet18 (mean DICE-Score of 94.10% vs. 91.78%) and exhibits the highest metric scores among all foundation models. In general, ViTs show competitive results to ResNet backbones in this case. It is noteworthy that the performance of ViT-S models not consistently surpasses that of their ViT-T counterparts, as measured by the average DICE-score. However, an increase in

Table 2. Body region classification results obtained with 5-fold cross-validation.

Backbone	Augmentation	Accuracy $\mu \pm \sigma$ (%)	Precision $\mu \pm \sigma$ (%)	Recall $\mu \pm \sigma$ (%)	F1-Score $\mu \pm \sigma$ (%)
ResNet18	Version A	73.78 ± 0.84	60.85 ± 2.82	73.82 ± 0.85	64.93 ± 0.92
ResNet50	Version A	75.00 ± 0.48	59.49 ± 0.38	75.05 ± 0.48	**65.95 ± 0.44**
ResNet50	Version B	73.16 ± 0.34	58.51 ± 0.37	73.19 ± 0.34	64.17 ± 0.33
ViT-T-16	Version A	97.26 ± 0.40	97.31 ± 0.42	97.27 ± 0.40	97.24 ± 0.41
ViT-T-8	Version A	97.26 ± 0.71	97.32 ± 0.68	97.27 ± 0.70	97.24 ± 0.72
ViT-S-16	Version A	97.67 ± 0.50	97.72 ± 0.51	97.68 ± 0.50	97.65 ± 0.51
ViT-S-16	Version B	97.50 ± 0.54	97.52 ± 0.55	97.51 ± 0.54	97.46 ± 0.56
ViT-S-8	Version A	97.80 ± 0.59	97.85 ± 0.57	97.80 ± 0.59	**97.79 ± 0.60**

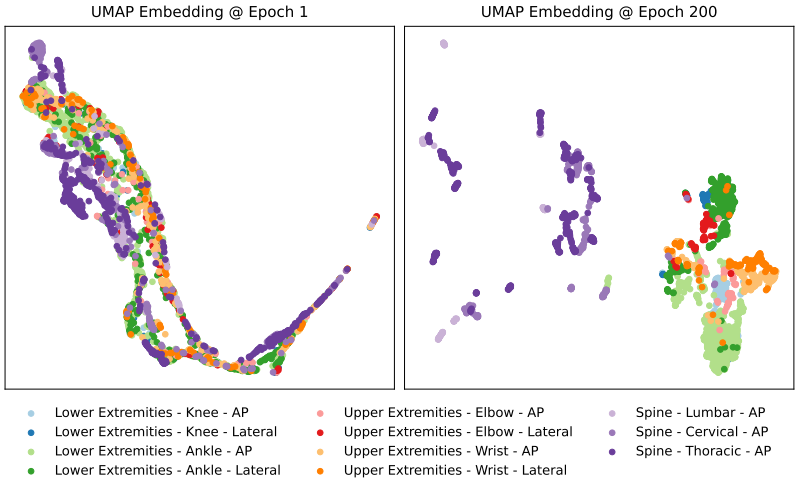

Fig. 2. UMAP projections for an untrained (left) and trained (right) ResNet50 backbone. Points are colored by body region class labels, which are not used for clustering.

resolution, achieved by reducing the patch size from 16x16 to 8x8 pixels, is found to slightly enhance the segmentation performance. Similar to the previous task, feature extractors from version B do not exhibit significant improvements or degradations compared to version A models. For a visual interpretation, Figure 3 provides exemplary segmentation mask predictions of different DAX backbones from version A.

Screw Object Detection. Table 4 summarizes the results of the third downstream task in terms of average precision (AP) and average recall (AR) according to different OKS thresholds. Generally, smaller thresholds result in higher performance scores. In this evaluation, we mainly focus on the mean AP (mAP) and mean AR (mAR) metrics, which are computed as the average values for varying OKS thresholds (0.50 to 0.95).

Table 3. Metal implant segmentation results obtained with 5-fold cross-validation.

Backbone	Augmentation	Accuracy $\mu \pm \sigma$ (%)	Precision $\mu \pm \sigma$ (%)	Recall $\mu \pm \sigma$ (%)	DICE-Score $\mu \pm \sigma$ (%)
ResNet18	Version A	99.39 ± 0.07	91.57 ± 0.99	92.55 ± 0.80	91.78 ± 0.72
ResNet50	Version A	99.57 ± 0.08	94.26 ± 1.03	94.17 ± 0.54	**94.10 ± 0.75**
ResNet50	Version B	99.46 ± 0.07	93.53 ± 0.63	93.80 ± 0.69	93.38 ± 0.61
ViT-T-16	Version A	99.27 ± 0.12	90.09 ± 0.78	90.80 ± 1.33	90.12 ± 0.69
ViT-T-8	Version A	99.41 ± 0.04	91.67 ± 1.41	93.13 ± 1.13	**92.09 ± 0.73**
ViT-S-16	Version A	99.28 ± 0.08	91.03 ± 0.81	91.29 ± 1.53	90.76 ± 1.20
ViT-S-16	Version B	99.34 ± 0.03	91.66 ± 0.60	91.39 ± 0.54	91.32 ± 0.48
ViT-S-8	Version A	99.26 ± 0.07	92.18 ± 1.08	91.23 ± 1.64	91.15 ± 1.06

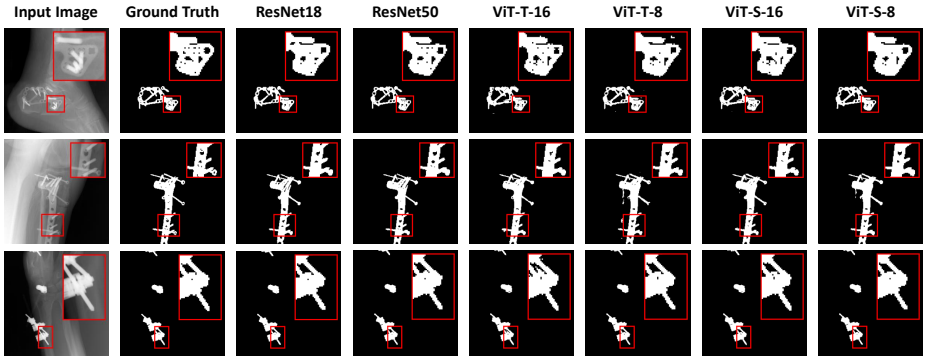

Fig. 3. Qualitative segmentation mask predictions of different backbone architectures pre-trained with DAX for three randomly selected samples of the test dataset.

All foundation models appear to overestimate screw objects, as indicated by higher mAR vs. mAP scores. In version A, ResNet18 and ResNet50 reveal similar performance, while ViT backbones exhibit inferior outcomes compared to ResNet architectures in general. Interestingly, ViT-S-8 achieves better results than ViT-T-8, whereas this relation is not observed in ViTs with 16x16 patches. Moreover, a reduction in patch size leads to enhanced detection performance exclusively for ViT-S, but not for ViT-T models. The backbones trained with version B show diminished mAP and equivalent mAR values, suggesting that there are no significant advantages over version A.

4 Discussion and Conclusion

Our findings reveal substantial differences between different DAX foundation models. ViTs achieve higher classification accuracy, while ResNet backbones perform better on segmentation as well as detection tasks. This contrast likely arises because the attention mechanism enables ViTs to capture global context [14,19], whereas ResNet architectures are strong in hierarchical feature extraction and spatial localization [8,15].

Table 4. Quantitative results for the screw object detection task.

Backbone	Augmentation	Average Precision (%)			Average Recall (%)		
		0.50:0.95	0.75	0.50	0.50:0.95	0.75	0.50
ResNet18	Version A	88.6	88.1	94.3	**91.7**	91.4	95.6
ResNet50	Version A	**89.0**	89.0	94.0	91.6	91.8	95.4
ResNet50	Version B	87.5	87.5	93.6	91.5	91.9	95.4
ViT-T-16	Version A	**86.1**	87.5	94.1	89.5	90.3	95.3
ViT-T-8	Version A	81.9	82.5	90.4	88.2	88.9	93.9
ViT-S-16	Version A	83.6	84.2	93.7	88.7	89.0	95.3
ViT-S-16	Version B	82.3	83.3	91.0	88.6	89.2	94.8
ViT-S-8	Version A	84.6	85.5	92.8	**89.9**	90.7	95.3

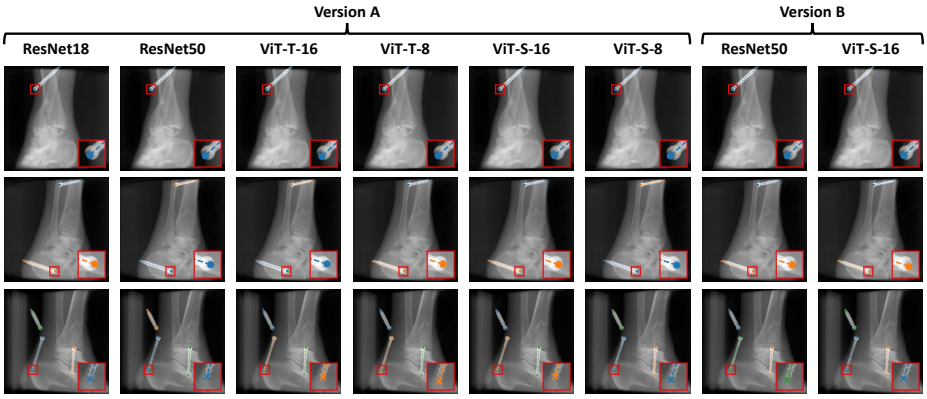

Fig. 4. Qualitative screw keypoint predictions of different backbone architectures pre-trained with DAX for three randomly selected samples of the test dataset.

Concerning data augmentation, models trained with version B exhibit either slightly diminished or comparable performance metrics as those from version A. Consequently, the latter version should be regarded as the preferred option. Our hypothesis is that the inherent diversity of the pre-training dataset may mitigate the impact of additional augmentation, despite the potential benefits of such operations in consideration of the underlying domain [7].

We primarily associate DAX with intraoperative X-ray imaging since most of the pre-training data originates from that domain. Therefore, the provided checkpoints are intended for use with intraoperative data only. However, the method itself could be applied to diagnostic or presurgical contexts as well.

Finally, our study is also subject to several limitations. First, the pre-training dataset is imbalanced with respect to several attributes, including age, gender, and body regions. Second, the evaluation is constrained to only three downstream tasks, which may not encompass the full range of clinical applications. Third, while frozen backbones offer computational efficiency and facilitate clearer interpretation of their feature extraction capabilities, this approach might restrict the overall model performance on specific downstream tasks.

In conclusion, we introduce DAX to demonstrate the effectiveness of self-supervised learning for pre-training foundation models in intraoperative X-ray imaging, thus supporting previous findings on the potential of such methods [23]. The DAX framework not only facilitates the development of novel, task-specific models but also effectively mitigates the challenges associated with the scarcity of labeled data. In the future, we will conduct experiments with further data augmentation techniques and a more extensive, balanced pre-training dataset.

Disclosure of Interests. As indicated by the affiliations, some authors were employees of Siemens Healthineers AG at the time of conducting this research. All other authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alam, I.S., Steinberg, I., Vermesh, O., van den Berg, N.S., Rosenthal, E.L., van Dam, G.M., Ntziachristos, V., Gambhir, S.S., Hernot, S., Rogalla, S.: Emerging Intraoperative Imaging Modalities to Improve Surgical Precision. Molecular Imaging and Biology **20**(5), 705–715 (Oct 2018)
2. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., Khan, F.S.: Foundation Models Defining a New Era in Vision: a Survey and Outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–20 (2025)
3. Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., Merhof, D.: Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision (Oct 2023). https://doi.org/10.48550/arXiv.2310.18689
4. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640. IEEE, Montreal, QC, Canada (Oct 2021)
5. Chan, H.P., Samala, R.K., Hadjiiski, L.M., Zhou, C.: Deep Learning in Medical Image Analysis. In: Deep Learning in Medical Image Analysis, vol. 1213, pp. 3–21. Springer International Publishing, Cham (2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021)
7. Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review **56**(11), 12561–12605 (Nov 2023)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K.: Are Natural Domain Foundation Models Useful for Medical Image Classification? In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7619–7628 (2024)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., others: Chexpert: A large chest radiograph

dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019), issue: 01
11. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (Dec 2019)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., others: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
13. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
14. Liu, Z., Qian, S., Xia, C., Wang, C.: Are transformer-based models more robust than CNN-based models? Neural Networks **172**, 106091 (Apr 2024)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
16. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024), publisher: Nature Publishing Group UK London
17. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction (Sep 2020). https://doi.org/10.48550/arXiv.1802.03426
18. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research (2024)
19. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do Vision Transformers See Like Convolutional Neural Networks? In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 12116–12128. Curran Associates, Inc. (2021)
20. Roh, B., Shin, J., Shin, W., Kim, S.: Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity. In: International Conference on Learning Representations (2022)
21. Saraf, V., Chavan, P., Jadhav, A.: Deep Learning Challenges in Medical Imaging. In: Vasudevan, H., Michalas, A., Shekokar, N., Narvekar, M. (eds.) Advanced Computing Technologies and Applications. pp. 293–301. Springer Singapore, Singapore (2020)
22. Shakouri, M., Iranmanesh, F., Eftekhari, M.: DINO-CXR: A Self Supervised Method Based on Vision Transformer for Chest X-Ray Classification. In: Bebis, G., Ghiasi, G., Fang, Y., Sharf, A., Dong, Y., Weaver, C., Leo, Z., LaViola Jr., J.J., Kohli, L. (eds.) Advances in Visual Computing. pp. 320–331. Springer Nature Switzerland, Cham (2023)
23. Shurrab, S., Duwairi, R.: Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Computer Science **8**, e1045 (Jul 2022)

24. Suetens, P.: Fundamentals of medical imaging. Cambridge university press (2017)
25. Unberath, M., Zaech, J.N., Gao, C., Bier, B., Goldmann, F., Lee, S.C., Fotouhi, J., Taylor, R., Armand, M., Navab, N.: Enabling Machine Learning in X-ray-based Procedures via Realistic Simulation of Image Formation. International journal of computer assisted radiology and surgery (IJCARS) (2019), publisher: Springer
26. Unberath, M., Zaech, J.N., Lee, S.C., Bier, B., Fotouhi, J., Armand, M., Navab, N.: DeepDRR–A Catalyst for Machine Learning in Fluoroscopy-guided Procedures. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer (2018)
27. Zaffino, P., Moccia, S., De Momi, E., Spadea, M.F.: A Review on Advances in Intra-operative Imaging for Surgery and Therapy: Imagining the Operating Room of the Future. Annals of Biomedical Engineering **48**(8), 2171–2191 (Aug 2020)