General Methods Make Great Domain-specific Foundation Models: A Case-study on Fetal Ultrasound

Jakob Ambsdorf[1,6]* , Asbjørn Munk[1,6]* , Sebastian Llambias[1,6] ,
Anders N. Christensen[2] , Kamil Mikolaj[2] , Randall Balestriero[5] ,
Martin G. Tolsgaard[1,3,4] , Aasa Feragen[2] , and Mads Nielsen[1,6]

[1] University of Copenhagen, Copenhagen, Denmark
[2] Technical University of Denmark, Kgs Lyngby, Denmark
[3] Copenhagen Academy of Medical Education and Simulation (CAMES),
Rigshospitalet, Copenhagen, Denmark
[4] Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark
[5] Brown University, Providence, Rhode Island, USA
[6] Pioneer Centre for AI, Copenhagen, Denmark
{jaam,asmu}@di.ku.dk

Abstract. With access to large-scale, unlabeled medical datasets, researchers are confronted with two questions: Should they attempt to pretrain a custom foundation model on this medical data, or use transfer-learning from an existing generalist model? And, if a custom model is pretrained, are novel methods required? In this paper we explore these questions by conducting a case-study, in which we train a foundation model on a large regional fetal ultrasound dataset of 2M images. By selecting the well-established DINOv2 method for pretraining, we achieve state-of-the-art results on three fetal ultrasound datasets, covering data from different countries, classification, segmentation, and few-shot tasks. We compare against a series of models pretrained on natural images, ultrasound images, and supervised baselines. Our results demonstrate two key insights: (i) Pretraining on custom data is worth it, even if smaller models are trained on less data, as scaling in natural image pretraining does not translate to ultrasound performance. (ii) Well-tuned methods from computer vision are making it feasible to train custom foundation models for a given medical domain, requiring no hyperparameter tuning and little methodological adaptation. Given these findings, we argue that a bias towards methodological innovation should be avoided when developing domain specific foundation models under common computational resource constraints.

Keywords: Self-supervised learning · Fetal Ultrasound · Foundation Models.

---

* Equal contribution.
[1] Code available at: https://github.com/jakobamb/UltraDINO. Model weights are available given permission to access pretraining data.
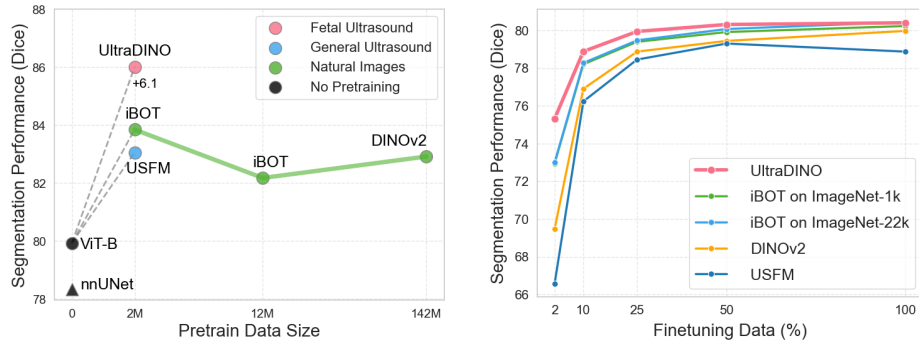
(a) Dice score on JNU-IFM few-shot segmentation trained on four patients, with up to ten images per patient.

(b) Few-shot performance vs. finetuning dataset size for pretrained models on segmenting fetal abdominal structures

Fig. 1: Segmentation performance depending on model and pretraining dataset.

1   Introduction

Recently, the medical imaging community experienced a surge in interest in foundation models, which, compared to purely supervised baselines, provide a strong starting point for training models that achieve better performance, require fewer labels, and are useful as general-purpose feature extractors [1,18,19,23]. Foundation models are pretrained using self-supervised learning (SSL) on large-scale datasets without additional labels, and can be finetuned on downstream tasks with only a few annotated samples [16]. Inspired by the success of these models in computer vision, various studies have investigated whether foundation models trained on natural images are useful for medical tasks as well [9,22]. However, the significant domain gap between natural images and the various medical imaging domains means that, while general visual features can be repurposed, the few-shot capability of these models remains limited, since semantics, such as patient anatomy, are not part of the data and are therefore not learned in pretraining [14]. This initiated research on medical data in an attempt to learn more semantic features, modeling patient anatomy and properties of the imaging modality. These approaches use either a large mixture of data from various modalities and domains [24], where individual medical areas may not be well represented, or medium-scale, modality-specific datasets [12,13,15,23]. A number of foundation models have since been proposed for medical ultrasound. USFM [13] is trained on a highly unbalanced dataset of approximately 2 million ultrasound images and uses a Masked-Image-Modeling (MIM) [20,8] approach with masking in both the spatial and the frequency domain. URFM [15] combines MIM on 1 million ultrasound images of different organs with knowledge distillation from a BiomedCLIP [26] model. UltraFedFM [12] is a concurrent work that uses a simple masked-autoencoder method in a multi-centre federated learning scenario on a diverse, 1 million image ultrasound dataset.
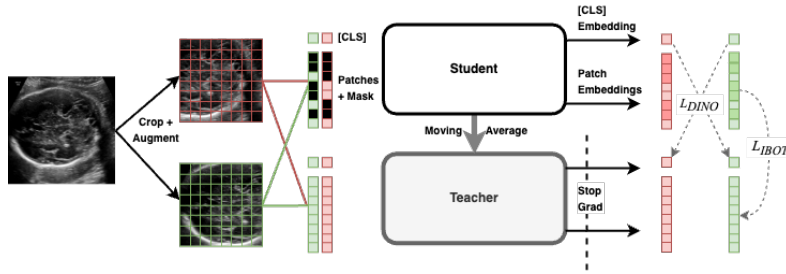
Fig. 2: UltraDINO uses the self-distillation method used by DINOv2, image adapted from [27]

Previous works on ultrasound foundation models used data from different anatomical regions, with fetal ultrasound not being well represented. At the same time, USFM and URFM introduced novel methodology, aimed at domain-specific improvements. In contrast, we investigate in this study how *well-established* methods from computer vision can be used to train medical foundation models, when abundant data from a *narrow domain* is available. We use the DINOv2 method to train UltraDINO, a foundation model for fetal ultrasound. Ultra-DINO is trained *exclusively* on a regional fetal ultrasound dataset, containing 2 million high-quality, clinical images. We evaluate the effectiveness of Ultra-DINO on three fetal ultrasound datasets from Spain, Brazil, and China, with tasks spanning standard-plane classification, coarse and fine anatomical segmentation, and various few-shot settings. Through our experiments, we distill two key insights:

1. It is worth the effort to pretrain on custom data, even if smaller models are trained on less data, as scaling in natural image pretraining does not translate to ultrasound.
2. Well-tuned methods from computer vision are making it feasible to train custom foundation models for a given medical domain, requiring no hyper-parameter tuning and little methodological adaptation.

Given these findings, we argue that a bias towards methodological innovation on the pretraining method can be detrimental, as tuning novel pretraining methods needs extensive hyperparameter optimization, requiring large compute resources not commonly available for medical imaging groups operating in specialized domains. In contrast, this study demonstrates that directly applying well-established pretraining methods to a sufficiently large, domain-specific dataset reaches state-of-the-art results for fetal ultrasound and generalizes well across tasks and countries.

2 Pretraining

Method. UltraDINO uses the self-distillation [6] method of iBOT [27] and DINO [3], which was later refined by DINOv2 [19], as shown in Figure 2. The

method is based on the following idea: Given two image encoders, called student and teacher, and two views of the same image (with different crops, augmentations, or maskings), the student attempts to predict both global (`[CLS]`-token) and local (masked patch-tokens) embeddings of the teacher. To prevent collapse, regularizations are applied, and the teacher is constructed as an exponential moving average of the student's parameters [6].

Data. Our pretraining dataset, denoted **FUS2M**, was collected the Danish national fetal ultrasound database, involving data from pregnant women who participated in prenatal ultrasound screening program in three out of five Danish regions between January 1st 2008 and June 30th 2018. The screening program consists of two ultrasound examinations, captured in the first and second trimester. We retrieved a random sample of 2 million ultrasound images with approvals from The Danish Data Protection Agency (Protocol No. P-2019-310) and by The Danish Patient Safety Authority (Protocol No. 3-3031-2915/1) from regional servers, excluding around 10.000 images that were used in previous studies. The fetal ultrasound examinations were performed using GE Logiq 7 (GE Healthcare, Milwaukee, WI, USA) or Voluson E6, E8, and E10 machines (GE Healthcare, Zipf, Austria).

Configuration. Our model uses ViT-B and ViT-S vision transformer architectures [5] with a patch size of 16, pretrained from scratch on FUS2M over 40 epochs, with a batch size of 512. All models were trained on two H100 80GB GPUs.

3   Experiments

We evaluate the pretrained models and non-pretrained baselines on one classification and two segmentation tasks. This includes evaluation of pretrained models using linear probing on classification, as well as full finetuning and comparison to non-pretrained baselines. We further evaluate the few-shot performance of models in different configurations. We report average results on the test set over three random seeds for full dataset experiments, and over five training splits in the few-shot settings. None of the evaluation datasets are included in FUS2M.

3.1   Baselines

We compare UltraDINO to the previously published general ultrasound foundation model USFM, as well as iBOT and DINOv2 which are trained on natural images. Since USFM did not publish code to reproduce their pretraining, we were unable to reproduce the method on FUS2M, and instead limit our evaluation to the published pretrained weights. USFM is finetuned using the recipes provided by the authors, since this yielded better results for the USFM model than

Table 1: **Performance comparison of models across classification and segmentation tasks.** Results are averaged and reported with standard error of the mean over three folds for full finetuning and five folds for few-shot learning. UltraDINO achieves superior few-shot learning and classification. JNU-IFM segmentation shows the largest performance gap. Linear probing suggests that UltraDINO embeddings are nearly linearly separable for standard planes.

	Method (Backbone)	PT Data	Segmentation (Dice↑)			Classification (F1↑)	
			JNU ($p=4$)	FASS ($n=20$)	FASS ($n=942$)	Fine-tune	LP
Pretrained	UltraDINO (ViT-B/16)	2 M FUS	**86.01**±0.74	**74.24**±0.38	**79.51**±0.11	94.54±0.16	**95.06**
	UltraDINO (ViT-S/16)	2 M FUS	85.58±0.76	73.29±0.42	79.14±0.02	**94.74**±0.24	94.17
	USFM (ViT-B/16)	2 M US	83.05±1.06	66.48±0.27	78.86±0.19	93.94±0.15	42.78
	DINOv2 (ViT-B/14 Dist.)	142 M Nat.	83.55±0.88	69.11±0.24	79.26±0.10	94.00±0.08	76.61
	DINOv2 (ViT-S/14 Dist.)	142 M Nat.	82.84±0.83	68.63±0.22	78.62±0.24	93.59±0.06	74.95
	iBOT (ViT-B/16)	12.5 M Nat.	82.17±1.37	72.01±0.38	79.24±0.14	94.16±0.07	83.01
	iBOT (ViT-B/16)	1 M Nat.	83.84±1.18	71.94±0.28	78.91±0.25	94.02±0.08	85.24
	iBOT (ViT-S/16)	1 M Nat.	81.94±1.30	72.23±0.21	79.09±0.16	94.01±0.06	83.98
From Scratch	nnUNet (UNet)	–	78.33±0.00	64.93±0.01	76.55±0.00	–	–
	ViT-S/16	–	79.59±1.18	62.54±0.12	77.95±0.18	85.08±0.40	–
	ViT-B/16	–	79.92±1.11	63.28±0.09	78.53±0.08	84.79±0.24	–
	MultiResUNet[10]	–	–	–	71.19±0.00	–	–
	UNet[21]	–	–	–	72.65±0.00	–	–
	UNetR[7]	–	–	–	69.96±0.00	–	–

using the configuration used for UltraDINO. For iBOT and DINOv2, we use official model checkpoints, which are trained on natural image datasets of different sizes, and on ViT-S and ViT-B models without pretraining. The iBOT models are trained on ImageNet-1K and ImageNet-22K consisting of 1.3M and 12.5M natural images respectively. DINOv2 is trained on LVD-142M consisting of 142M natural images. The DINOv2 checkpoints are distilled from a ViT-g model with 1.1B parameters. All models use patch size 16, except DINOv2 which is only available with patch size 14. Lastly, we evaluate non-pretrained performance of nnUNet [11], as well as other UNet variants outside of the nnUnet framework for segmentation performance. As the best performing supervised baseline, only nnUNet is also evaluated for few-shot.

3.2 Segmentation

Segmentation models are trained by attaching a segmentation decoder head to the model. We follow [13,27,19] and use UperNet [25] as decoder and fine-tune both encoder and decoder. For fair comparison, all models are trained on 224x224 resolution images. We use the dice similarity score to measure the segmentation performance. To gauge the segmentation performance of the foundation models, we use the FASS dataset and the JNU-IFM dataset.

<u>FASS</u> consists of 1,588 ultrasound images from 169 term pregnant women captured at a hospital in Santa Catarina, Brazil [4]. It includes expert-annotated

segmentations of four anatomical structures: the fetal abdominal aorta artery, intrahepatic umbilical vein, stomach, and liver. We use three configurations, in which the same 20% of the images are reserved for testing: (1) *full dataset* in a 80/20 train/val split, (2) *few-shot* with 20 images for training (approximately 2% of the dataset) and the remaining for validation, (3) *label-efficiency* using 10%, 25% and 50% of the dataset for training and the remaining for validation.

JNU-IFM contains 6,224 images from 78 ultrasound-videos, recorded from 51 patients at the Intelligent Fetal Monitoring Lab of Jinan University, China [17]. It features a relatively simple segmentation task of deliniating the pubic symphysis and fetal head. We limit our experiments to a few-shot setting where we train on only the data from 4 patients, with a maximum of 10 images sampled per patient. Approximately half of the dataset is held out for testing, and validation is performed on all images not used for training. We filter out all images on which at least one anatomical structure is not visible.

3.3   Classification

Classification models are trained on the Fetal Planes dataset [2] using both full end-to-end fine-tuning and linear probing, using the F1 score to measure classification performance. The Fetal Planes dataset is a standard plane classification dataset containing 12,400 images from 1792 patients and six classes, recorded at Hospital Clinic and Hospital Sant Joan de Deu, Barcelona, Spain [2]. Standard planes are specific views of the fetal anatomy as defined by clinical criteria. We test on the provided split of 5272 images and divide the remaining training data into an 80/20 split of train and validation.

Linear Probing. We evaluated our pretrained model using linear probing by freezing its feature extractor and training a simple linear classifier on top of the learned representations. This allowed us to assess the quality of the pretrained features without fine-tuning the entire model. We follow the evaluation approach of [19] by training a series of linear classifiers with different configurations on the train split while keeping the backbone fixed, and evaluate the best linear classifier according to validaiton accuracy on the test set.

4   Results and Discussion

Table 1 provides an overview of the results. In general, UltraDINO models, both on small and base-size ViTs outperform all baselines in few-shot and classification tasks. On the segmentation tasks, the largest separation between UltraDINO and other models can be observed on the few-shot version of the JNU-IFM dataset. This is noteworthy, as the pubic symphysis/fetal head are, while being obstetric images, not part of the fetal screening protocol and therefore out-of-domain for the models trained on the FUS2M dataset. On FASS *full*, all pretrained models reach similar performance, which could indicate other limiting factors, such as label noise and decoder precision. The *few-shot* configuration is more informative
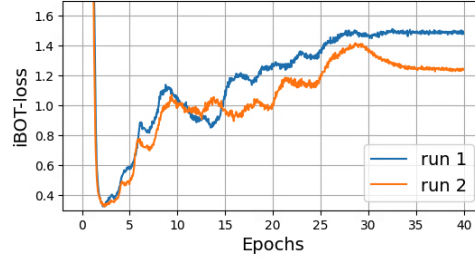
Fig. 3: iBOT-loss curves for two runs with the same configurations *and same seed* using a ViT-S on FUS2M are diverging considerably when using default non-deterministic algorithms.

for evaluating the embedding quality. Here, USFM is outperformed by all iBOT and DINOv2 baselines, while UltraDINO leads by 1-2 Dice points, depending on model size. The *label efficiency* over different fractions of the FASS training dataset is represented in Figure 1b and shows a similar trend, where UltraDINO reaches close to full performance on 25% of the training data. Interestingly, larger pretraining datasets do not improve the performance of the natural image foundation models, as can be observed in Figure 1a.

After finetuning on the FetalPlanes dataset, the ViT-S of UltraDINO performs slightly better than the ViT-B, with all baselines trailing by about 0.5% on F1. Linear probing reveals that UltraDINO embeddings are almost entirely linearly separable in the dataset classes. Surprisingly, all iBOT models perform very well and better than the DINOv2 models. USFM's performance on linear probing is poor, as it is common for masked auto-encoder-based models [8].

Overall, any pretraining method is beneficial and significantly improves upon the non-pretrained baselines. USFM is outperformed by the natural image foundation models in the majority of cases, while UltraDINO models lead by a margin on few-shot and linear probing experiments. The results demonstrate that a well-tuned natural image foundation model outperforms domain-specific foundation models if the pretraining dataset is insufficient for the target domain and non-state-of-the-art pretraining methods are used, while the application of leading pretraining methods to high-quality pretraining datasets consistently improve upon natural image baselines. Notably, the pretraining dataset is not regionally overlapping with the finetuning tasks, indicating strong generalization capability despite this domain-shift.

Tuning DINOv2 for fetal ultrasound. During this study, we attempted to tune the hyperparameters of DINOv2 to match the ultrasound domain. We conducted experiments on ViT-S models with an iteration time of about 8 hours (at 318.39 GFLOPs per forward pass), but found that reaching confident conclusions on shorter training runs is inhibited by the high stochasticity intrinsic to the pretraining method. We observe that through large differences in the iBOT
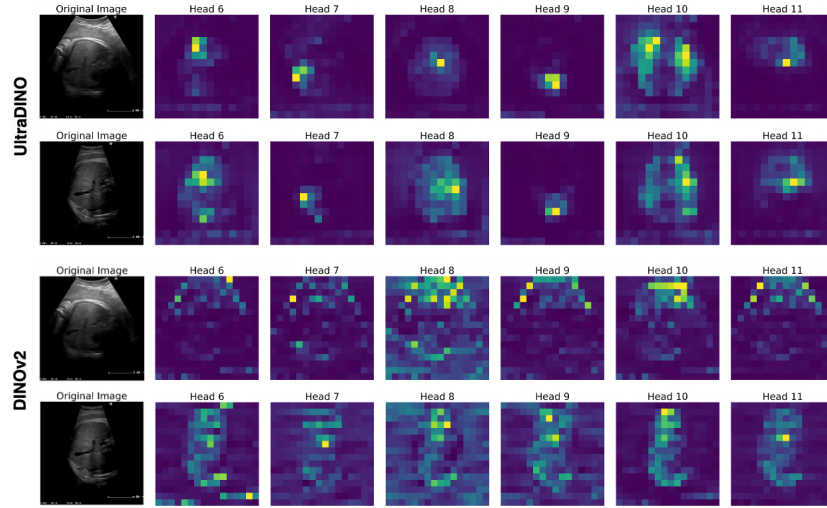
Fig. 4: Attention maps of DINOv2 and UltraDINO: UltraDINO attention heads display semantic specialization to anatomical structures across images.

loss for the same model configuration, even when using the same random seed when using PyTorch's standard and significantly faster non-deterministic algorithms (see Figure 3), adding another layer of complexity and computational cost for experimentation on the pretraining method.

Qualitative evaluation of attention maps. Figure 4 shows the attention maps of UltraDINO and DINOv2 ViT-B models on the FASS validation set. We observe that while DINOv2 models capture high contrast edges, there is little to no specialization across attention heads. In contrast, UltraDINO attention maps are semantic and tend to relate to the same anatomical structures across images.

Limitations. Our study has several limitations. Firstly, while we compare UltraDINO to USFM, DINOv2, and iBOT, we do not include other self-supervised learning (SSL) methods beyond self-distillation pretrained on natural or medical images. Secondly, the lack of available training code for USFM prevented us from reproducing its method on our dataset, limiting our analysis to the pretrained USFM weights provided by the authors. Further, while we evaluate on publicly available datasets from different countries, access to more diverse and standardized datasets would enable a more comprehensive assessment of generalization, especially for low-resource settings.

5   Conclusion

Through the case study of UltraDINO, we demonstrated using established methods that pretraining on a large-scale, high-quality dataset from a single medical

domain reaches state-of-the-art performance and consistently outperforms generalist models trained on natural images, even when trained on substantially larger datasets. Our results emphasize two key findings: (i) Domain-specific pretraining is highly beneficial, and even smaller models trained on less data outperform larger models pretrained on natural images. (ii) The choice of pretraining method matters: By choosing a well-established self-supervised learning method, it is possible to achieve state-of-the-art performance across multiple fetal ultrasound tasks with no additional hyperparameter tuning. Based on these findings, and despite the computational cost and the stochastic variance of pretraining, we recommend training custom medical foundation models if relevant high-quality training data is available. However, instead of focusing on novel methodological development, well-tuned and established methods from computer vision are highly effective and should form the basis for any further experimentation.

Disclosure of Interests. ANC, MT, AF, and MN hold shares in Prenaital ApS.

References

1. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
2. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F., Gratacós, E.: Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. Scientific Reports **10**(1), 10200 (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Correggio, K.S.D., Galluzzo, R.N., Santos, L.O., Barroso, F.S.M., Chaves, T.Z.L., Onofre, A.S.C., von Wangenheim, A.: Fetal abdominal structures segmentation dataset using ultrasonic images (Oct 2023). `https://doi.org/10.17632/4gcpm9dsc3.1`, `https://data.mendeley.com/datasets/4gcpm9dsc3/1`
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)

7. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
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: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 7634–7643 (2024)
10. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)
11. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
12. Jiang, Y., Feng, C.M., Ren, J., Wei, J., Zhang, Z., Hu, Y., Liu, Y., Sun, R., Tang, X., Du, J., et al.: Privacy-preserving federated foundation model for generalist ultrasound artificial intelligence. arXiv preprint arXiv:2411.16380 (2024)
13. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
14. Juodelyte, D., Lu, Y., Jiménez-Sánchez, A., Bottazzi, S., Ferrante, E., Cheplygina, V.: Source matters: Source dataset impact on model robustness in medical imaging. In: International Workshop on Applications of Medical AI. pp. 105–115. Springer (2024)
15. Kang, Q., Lao, Q., Gao, J., Bao, W., Lu, Q., Li, K.: Urfm: a general ultrasound representation foundation model for advancing ultrasound image diagnosis. iScience (2024)
16. Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (2022)
17. Lu, Y., Zhou, M., Zhi, D., Zhou, M., Jiang, X., Qiu, R., Ou, Z., Wang, H., Qiu, D., Zhong, M., et al.: The jnu-ifm dataset for segmenting pubic symphysis-fetal head. Data in brief **41**, 107904 (2022)
18. Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. (2024)
20. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
22. Song, X., Xu, X., Yan, P.: General purpose image encoder dinov2 for medical image registration. arXiv preprint arXiv:2402.15687 (2024)

23. Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Maerkisch, L., Jäger, P.F., Maier-Hein, K.: Revisiting mae pre-training for 3d medical image segmentation. arXiv preprint arXiv:2410.23132 (2024)
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463 (2023)
25. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: European Conference on Computer Vision. Springer (2018)
26. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
27. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)