# Towards General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 on Radiology Benchmarks

Mohammed Baharoon[a], Waseem Qureshi[d], Jiahong Ouyang[b], Yanwu Xu[c], Kilian Phol[b], Abdulrhman Aljouie[e], Wei Peng[b,*]

[a] *Fatima Fellowship*

[b] *Stanford Medicine, Stanford University, Stanford, CA 94305*

[c] *Boston University, Boston, MA 02215*

[d] *Oxford Cancer, University of Oxford, Oxford, OX3 7LA*

[e] *King Abdullah International Medical Research Center*

## Abstract

The integration of deep learning systems into the medical domain has been hindered by the resource-intensive process of data annotation and the inability of these systems to generalize to different data distributions. Foundation models, which are models pre-trained on large datasets, have emerged as a solution to reduce reliance on annotated data and enhance model generalizability and robustness. DINOv2, an open-source foundation model pre-trained with self-supervised learning on 142 million curated natural images, excels in extracting general-purpose visual representations, exhibiting promising capabilities across various vision tasks. Nevertheless, a critical question remains unanswered regarding DINOv2's adaptability to radiological imaging, and the clarity on whether its features are sufficiently general to benefit radiology image analysis is yet to be established. Therefore, this study comprehensively evaluates DINOv2 for radiology, conducting over 100 experiments across diverse modalities (X-ray, CT, and MRI). Tasks include disease classification and organ segmentation on both 2D and 3D images, evaluated under different settings like kNN, few-shot learning, linear-probing, end-to-end fine-tuning, and parameter-efficient fine-tuning, to measure the effectiveness and generalizability of the DINOv2 feature embeddings. Comparative

[*] Corresponding author
*Email address:* wepeng@stanford.edu (Wei Peng)

analyses with established medical image analysis models, U-Net and TransUnet for segmentation, and CNN and ViT models pre-trained via supervised, weakly supervised, and self-supervised learning for classification, reveal DINOv2's superior performance in segmentation tasks and competitive results in disease classification. The findings contribute insights to potential avenues for optimizing pre-training strategies for medical imaging and enhancing the broader understanding of DINOv2's role in bridging the gap between natural and radiological image analysis.

*Keywords:* Representation Learning, Self-Supervised Learning, Foundation Models, Computer Vision, Medical Image Analysis.

## 1. Introduction

The field of computer vision has recently seen a rise in interest for general-purpose models that are optimized to function across different tasks and domains [1, 2, 3, 4]. These models, grouped under the term "Foundation Models" (FMs), usually contain parameters ranging from hundreds of millions to tens of billions and are trained on large datasets, on the order of tens of millions. As a result of this large-scale training, these FMs often achieve state-of-the-art (SoTA) results and impressive zero-shot and few-shot performance and generalizability [3, 4]. For these reasons, foundation models have gained traction in deep learning-based medical image analysis research [5, 6, 7, 8, 9], as it holds promise for reducing the reliance on the expensive process of annotating medical data and towards the goal of building generalist medical artificial intelligence systems that can function across a variety of tasks [10].

### 1.1. What Are Foundation Models?

The term foundation model is an umbrella term that covers a wide range of different models. In the most general sense, foundation models are large models trained on large datasets and can generalize across tasks and/or domains [11]. To make the term more useful in our analysis, however, we group FMs using two methods of categorization. First, we divide FMs depending on their training paradigm into three groups: self-supervised, weakly-supervised, and supervised foundation models. Weakly-supervised and supervised foundation models require correspondence in the training data. In these paradigms, the training data is required to be available in pairs: an X-ray examination and a corresponding interpretation, diagnosis, or segmentation mask for example. Models like OpenAI's CLIP [2] and Meta's Segment Anything Model (SAM) [4] fall under this category. Self-supervised foundation models, on the other hand, require one input data to train (an image, text, etc.) Meta's DINOv2 [3] and Google's Universal Speech Model (USM) [12] belong to this category.

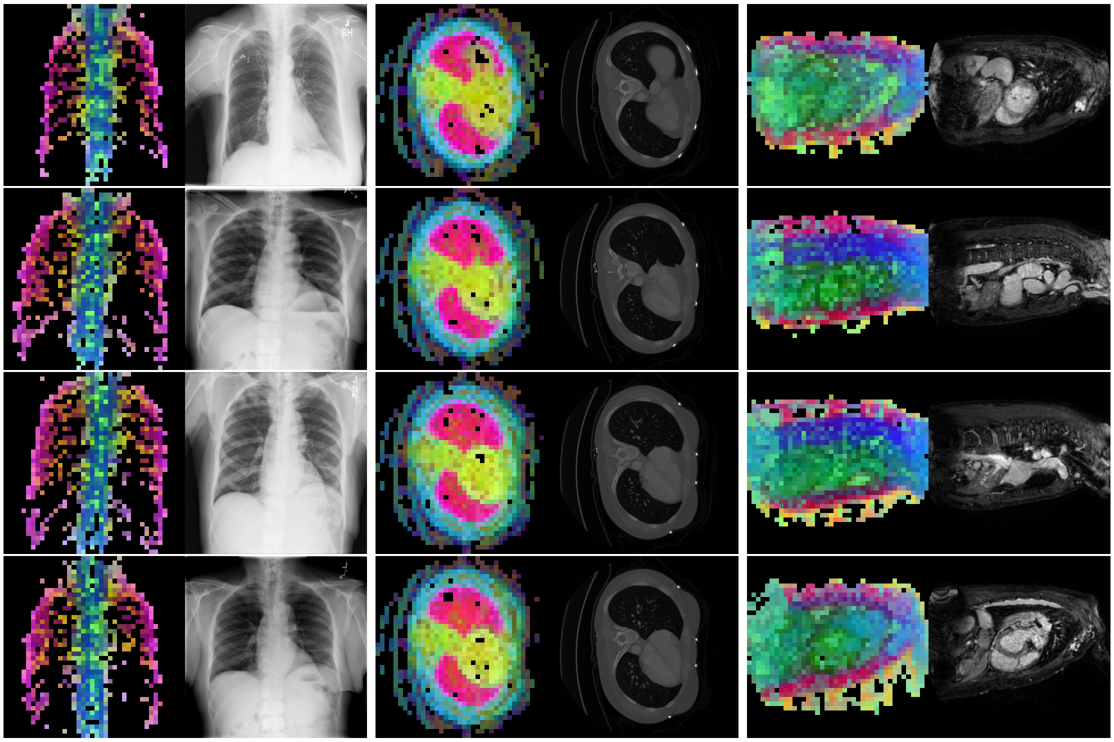
Additionally, we categorize FMs into two groups depending on the generalizability of their produced representations: general purpose (also called task-agnostic), and task-specific FMs. General purpose foundation models produce features that generalize across more than one task, segmentation and classification for example, while task-specific models specialize on only one task. DINOv2 [3] and USM [12] fall under the former category while SAM [4] is under the latter.

Because of the less restrictive training dependencies for self-supervised FMs, we believe that they are a promising option to explore for future research especially for the medical imaging domain, since there is an abundance of public medical image data without corresponding descriptions or diagnoses, either due to patient privacy barriers, data collection overhead, or other issues. Moreover, FMs that produce representations that can be used across a variety of tasks are desired because it is usually a signal of model robustness. As a result, we focus our attention on self-supervised general-purpose FMs for medical image analysis. Specifically, we adopt Meta's DINOv2 [3] model, a publicly available general-purpose vision foundation model that can extract robust representations across different vision tasks, for experimentation on a wide range of medical disease classification and organ segmentation benchmarks, across different radiological exams and under different evaluation settings.

### 1.2. What is DINOv2?

DINOv2 is a successor of DINO [13] and constitutes both a self-supervised pre-training method based on DINO and iBOT [14], and a collection of models pre-trained using that method. It was released to the public by Meta in April 2023 and promises robust representations that enable general-purpose functionality with visual-only data [3]. The released models were pre-trained on a dataset of 142 million carefully curated natural images, called LVD-142M. Roughly 100 million of these are images that are similar to ImageNet, curated from calculating similarly of web-scarpered images with ImageNet21k dataset [15]. The remaining images were retrieved based on their similarity to Caltech 101 [16], ADE20k [17], and Google Landmarks v2 [18], among others. The models achieve competitive performance on classification, segmentation, depth estimation, and image retrieval tasks across both image and video benchmarks. Moreover, because of the adopted discriminative self-distillation approach, the DINOv2 performs well “out-of-the-box,” without the need to fine-tune the encoder. The ViT-g/14 version of the model achieves an accuracy of 86.5% and 83.5% on Linear-probing and kNN evaluations, respectively, on ImageNet-1k, outperforming other weakly-supervised and self-supervised methods. This capability to perform well out-of-the-box is appealing, especially in the medical domain, as it implies competitive performance even in low-data and low-computation settings.

**Figure 1: PCA component visualization.** Following [3], the PCA is computed between patches of images that are in the same column, and the first 3 components are shown. This is done for X-ray, CT, and MRI scans. Thresholding is used on the first PCA component to remove the background.



### 1.3. Contribution

In this paper, we set to work towards vision-only general-purpose foundation models for the medical domain by adopting DINOv2 for disease classification and organ segmentation benchmarks. We perform comprehensive evaluations of DINOv2 across various scenarios for multiple radiology modalities, exploring both low (few-shot) and high data settings spanning X-ray, CT, and MRI benchmarks.

We evaluate DINOv2 on both disease classification and organ segmentation benchmarks. For disease classification tasks, we evaluate the model on kNN, linear-probing, few-shot learning, parameter-efficient fine-tuning, and end-to-end fine-tuning scenarios. For organ segmentation tasks we compare lightweight and non-lightweight decoders, while we keep the DINOv2 backbone frozen. As far as we know, there is no comprehensive analysis of DINOv2 on medical benchmarks that evaluates the model across disease classification and organ segmentation tasks and on different radiological examinations. Our contributions can be summarized as follows:

- We evaluate all versions of DINOv2 ViT models (small, base, large, and giant) on X-ray, CT, and MRI disease classification benchmarks, under kNN, linear-probing, few-shot learning, parameter-efficient fine-tuning, and end-to-end fine-tuning. We also provide a comparison with other supervised, self-supervised, and weakly-supervised models.

- We test the performance of DINOv2 for organ segmentation tasks across X-ray, CT, and MRI modalities, by freezing the DINOv2 encoder and attaching a linear and U-Net decoder. We compare the results to U-Net and TransUnet models trained end-to-end.

- We compare the ability of DINOv2 to generalize across tasks with supervised classification and segmentation models. We specifically compare the results of DINOv2 to a supervised pre-trained classification model when evaluated on organ segmentation tasks and DINOv2 with the Segment Anything Model when evaluated on disease classification tasks.

- We employ parameter-efficient fine-tuning methods like LoRA and BitFit to tune the DINOv2 ViT-g/14 model on large X-ray classification datasets like NIH Chest X-ray and CheXpert, and provide a performance and efficiency comparison of PEFT with end-to-end fine-tuning and linear-probing. We conclude that using PEFT yields performance that is competitive with end-to-end fine-tuning using less than 1% of the total model parameters.

- We provide a large-scale radiology benchmark with multiple image modalities, and evaluation pipelines that can be used for general-purpose FM testing.

## 2. Related Works

Other works have explored applying weakly-supervised, self-supervised, and supervised foundation models and learning paradigms for medical image analysis [19, 9, 20].

Previous works have explored weakly-supervised vision-language models for zero-shot classification, visual question answering, and image generation in the medical domain, achieving SoTA or competitive performance on radiology benchmarks [19, 21, 22]. Wang et al. [19] proposed MedCLIP, employing an extension of CLIP's contrastive pre-training method that utilizes unpaired examination-report samples, using the MIMIC-CXR [23] and CheXpert [24] datasets and outperforming previous SoTA on zero-shot and supervised classification on four radiology benchmarks. Yet, the performance of MedCLIP, along with other vision-language models, is limited by the availability of text-image data for training.

After the release of SAM, other segmentation-specific models were adopted for the medical domain [20]. Ma et al. [9] have developed MedSAM by adopting SAM to the medical domain, using a dataset of more than one million medical images across diverse segmentation tasks and examination modalities. MedSAM significantly outperforms SAM on medical tasks and achieves comparable performance to U-Net [25] models. Still, MedSAM is limited to medical segmentation and cannot be generalized to other medical image analysis tasks.

Self-supervised pre-training applied to medical datasets has recently achieved SoTA on CT and MRI segmentation benchmarks [26, 27]. Tang et al. [27] pre-trained a Swin UNETR [28] architecture with self-supervision on 5,050 CT volumes and outperformed previous SoTA on the BTCV [29] and MSD [30] competitions. However, to the best of our knowledge, there is still no self-supervised general-purpose FM for the medical domain that has been proven to achieve consistently competitive results across tasks and radiological examinations.

## 3. Methodology

In this section, we describe the motivation for adopting DINOv2 for this study and outline the settings under which we performed our experiments. The pre-processing and hyper-parameter tuning pipeline can be viewed in the GitHub repository at 6.

### 3.1. Motivation for Using DINOv2

There are many vision foundation models trained on natural images with supervised, self-supervised, and weakly-supervised learning including CLIP [2], SAM [4], and Florence [1, 31]. Additionally, there are many other vision foundation models trained on medical data that achieve SoTA on a selected number of competitions and benchmarks [26, 27]. However, we decided to employ DINOv2 in our analysis because of the robustness of its representations, achieving competitive performance in multiple downstream tasks, across vision modalities (image and video), and, most importantly, in out-of-the-box evaluations. The DINOv2 training paradigm was specifically designed to generate powerful representations on out-of-the-box kNN evaluations and outperforms many other weakly-supervised and self-supervised foundation models in kNN and linear-probing [3]. When comparing with foundation models trained on medical data, we chose DINOv2 because it was trained on a much larger dataset compared to the medical-specific models (142 million 2D images vs 5,050 [27] or 10,000 3D volumes [26]) and has a much larger parameter count (61.98M for [27] and 1,100M for DINOv2 [3]). We believe that, as of now, this is one of the best fit open-source FM for the purposes of this analysis.

### 3.2. Datasets

We evaluated DINOv2 on 8 public radiology benchmarks, spanning X-ray, CT, and MRI examinations (Exam.) for disease classification (CLS, 4 datasets) and organ segmentation (SEG, 4 datasets) tasks. A summary of the used datasets is shown in Table 1. "# Classes" and "# Images" describe the number of classes and the number of images employed in our analysis, respectively, and not the number in the original dataset. For some of the datasets shown, we only selected a sample of the entire available data. All the datasets used are publicly available. The data pre-processing pipeline is available in the GitHub repository in section 6. Some of the datasets used do not have a predefined test set. On these datasets, we used systemic sampling to divide the dataset into train, evaluation, and test subsets. The dataset splits are provided in the Google Drive link in 6.

### 3.3. Evaluation Settings

In the analysis, we focused mainly on the "out-of-the-box" performance of DINOv2, where we trained a classification or segmentation heads while keeping the backbone frozen. This resulted in preferable lightweight training that requires fewer labeled instances, computational resources,

**Table 1: The datasets used in our analysis.** For datasets that do not have a standardized test set, we chose the test set using systematic sampling. All the datasets and splits are public.

| Dataset | Exam. | Task | Labels | # Classes | # Images | Dim |
|---|---|---|---|---|---|---|
| NIH Chest X-ray [32] | X-ray | CLS | Thorax Diseases | 14 | 112,120 | 2D |
| CheXpert [24] | X-ray | CLS | Thorax Diseases | 5 | 161,792 | 2D |
| Montgomery County [33] | X-ray | SEG | Lung Masks | 3 | 138 | 2D |
| Shenzhen [33] | X-ray | SEG | Lung Masks | 2 | 566 | 2D |
| SARS-CoV-2 [34] | CT | CLS | COVID-19 Diagnosis | 2 | 4,173 | 3D |
| AMOS [35] | CT | SEG | Abdominal Organs Masks | 15 | 21,124 | 3D |
| MSD Heart [30] | MRI | SEG | Left Atrium Masks | 2 | 2,271 | 3D |
| Brain Tumor [36] | MRI | CLS | Tumor Types | 3 | 3,064 | 2D |

and training time. Additionally, we also performed end-to-end fine-tuning and parameter-efficient fine-tuning evaluations for performance comparison to this lightweight training paradigm.

We experimented with the original DINOv2 ViT-g/14 and the three smaller distilled versions (ViT-L/14, ViT-B/14, and ViT-S/14). Section 3.3.1 further details the evaluation settings for the classification experiments, while 3.3.2 does the same for segmentation settings.

### 3.3.1. Disease Classification

For disease classification evaluations, we compare the performance of all the DINOv2 models to other large supervised, weakly-supervised, and self-supervised CNN and ViT models. Table 2 shows the classification backbones used along with their size and pre-training settings.

We performed four main types of experiments: kNN, linear-probing, few-shot learning, and fine-tuning. (1) kNN was performed on the normalized features of the last backbone layer. (2) For Linear-probing, a single linear layer was attached on top of the backbone. (3) In few-shot learning, we trained a linear layer on top of frozen features. (4) For fine-tuning, we used both parameter-efficient fine-tuning methods like LoRA [37] and BitFit [38] and end-to-end fine-tuning.

When using ViT architectures, the linear layers take either the CLS token or the CLS token concatenated with the average of all patch tokens, depending on which method yielded higher performance in the validation set. For 3D CT volumes, the embeddings for all slices were averaged before being passed into the classification head.

**Table 2: Models used for classification.** Description of the classification backbones used along with their parameter count and pre-training settings.

| Method | Architecture | Dataset | # Images | # Params. |
|---|---|---|---|---|
| DINOv2 | ViT-g/14 | LVD-142M | 142M | 1,100M |
| | ViT-L/14 | LVD-142M | 142M | 300M |
| | ViT-B/14 | LVD-142M | 142M | 86M |
| | ViT-S/14 | LVD-142M | 142M | 21M |
| Supervised | ViT-L/16 | ImageNet21k | 14M | 300M |
| | VGG19 | ImageNet1k | 1.3M | 144M |
| | ResNet152 | ImageNet1k | 1.3M | 60M |
| | DenseNet201 | ImageNet1k | 1.3M | 20M |
| MAE | ViT-L/16 | ImageNet1k | 1.3M | 300M |
| CLIP | ViT-L/14 | WIT-400M | 400M | 300M |

### 3.3.2. Organ Segmentation

For organ segmentation evaluations, we compared using a frozen DINOv2 ViT-g/14 and DINOv2 ViT-L/14 encoder with attached decoder, to segmentation architectures that are commonly used in medical image analysis, like U-Net and TransUnet. Table 3 described the models used in the segmentation experiments.

**Table 3: Description of the segmentation architectures.** The parameter count given is when there is one output class.

| Method | Architecture | Trainable Params. (%) | Total Params. |
|---|---|---|---|
| Scratch | U-Net | 31M (100%) | 31M |
| | TransUnet | 324M (100%) | 324M |
| DINOv2 | ViT-L/14-Linear | 1,000 (< 1%) | 300M |
| | ViT-L/14-U-Net | 17M (5%) | 317M |
| | ViT-g/14-Linear | 1,500 (< 1%) | 1,100M |
| | ViT-g/14-U-Net | 38M (3%) | 1,138M |

We experimented with both a lightweight single linear layer decoder and a U-Net-based, hierarchical decoder. The U-Net decoder is made up of four blocks, where each block consists of one convolutional layer along with ReLU activation function and batch normalization. Skip connections were obtained from the previous four blocks of the transformer model and concatenated to the features at each U-Net layer, as is done in the classical U-Net architecture [25]. For processing 3D volumes, we segmented each slice independently.

## 4. Results

In this section, we report our results across the different evaluation settings, tasks, and radiological modalities. We evaluated the DINOv2 ViT-g/14 model and the smaller distilled versions. For disease classification tasks, we compare DINOv2 to other large supervised classification models that are commonly used as backbones for transfer learning in the medical domain, like ViT-L/16 [39] pre-trained on ImageNet21k, DenseNet201 [40], ResNet152 [41], and VGG19 [42] pre-trained on ImageNet1k. We also compare self-supervised MAE [43] and weakly-supervised CLIP [2] pre-trained ViTs. For organ segmentation, we compare a U-Net decoder on top of frozen DINOv2 features with a U-Net and TransUnet models trained end-to-end from scratch. We used the area under the operating receiver curve (AUROC) as a performance metric for classification tasks, and the average of the dice and jaccard scores as a metric for segmentation.

In section 4.1 we will first compare DINOv2 with commonly-used supervised, weakly-supervised, and self-supervised models on kNN, linear-probing, and fine-tuning settings for disease classification tasks. Moreover, we will compare using a frozen DINOv2 encoder with a U-Net decoder on organ segmentation datasets to U-Net and TransUnet models trained end-to-end from scratch. We also show a comparison between using a linear layer decoder and a U-Net, hierarchical decoder on top of the frozen DINOv2 ViT-L/14 features. Then, in section 4.2, we will explore the few-shot learning capability of DINOv2 on both disease classification and organ segmentation tasks. Specifically, we will compare the results the few-shot results of DINOv2 to supervised models on disease classification tasks and to U-Net and TransUnet models on organ segmentation tasks. In section 4.3, we perform parameter-efficient fine-tuning to compare the performance and efficiency of using PEFT with end-to-end fine-tuning and linear probing. In section 4.4, we analyze the cross-task generalizability of DINOv2 compared to other supervised models on disease classification and organ segmentation tasks. Finally, 4.5 shows qualitative results of DINOv2 features on X-ray, CT, MRI modalities, and organ segmentation results of linear and U-net decoders trained on top of frozen DINOv2 ViT-L/14 features.

### 4.1. Disease Classification and Organ Segmentation

Table 4 shows the linear-probing performance of all DINOv2 models compared to supervised learning methods on X-ray, CT, and MRI disease classification datasets. DINOv2 performs on par or slightly better compared to the other methods, and outperforms by a relatively larger margin commonly-used CNN supervised models on linear-probing.

Table 5 shows the kNN, linear-probing, and end-to-end fine-tuning results of DINOv2 compared to other supervised, weakly-supervised, and self-supervised methods on the NIH Chest X-ray and CheXpert datasets. DINOv2 outperforms the self-supervised and weakly-supervised methods in linear-probing and end-to-end fine-tuning. It also outperforms other supervised methods when fine-tuning on the NIH Chest X-ray dataset, but not CheXpert. Also important to note that DINOv2 under-performs on kNN evaluations compared to other methods, even though its features were designed to maximize kNN results. This can be explained by the domain shift between the pre-training natural images and medical images, making the out-of-the-box kNN evaluations more random.

Moreover, Table 6 shows a performance comparison between using a linear layer decoder and a U-Net decoder on top of frozen DINOv2 ViT-L/14 features on four organ segmentation datasets. The linear layer decoder evaluations are used to isolate the performance of the DINOv2 encoder, analogous to the purpose of kNN classification evaluations. The single linear layer performs surprisingly well on 3 out of 4 datasets. Figure 3 shows qualitative segmentation for both methods.

Finally, in Table 7 we compare U-Net decoders on top of frozen DINOv2 ViT-L/14 and ViT-g/14 features with U-Net and TransUnet models trained end-to-end from scratch. The results of all models are similar on the easier Montgomery Country, Shenzhen, and MSD Heart datasets, but DINOv2 outperforms the other models on the more difficult AMOS multi-organ segmentation dataset, even with a frozen encoder and less trainable parameters.

### 4.2. Few-shot Learning

We perform few-show learning for disease classification and organ segmentation tasks on X-ray images. On the left of Figure 2 is the performance of different backbones on the NIH Chest X-ray

**Table 4: Linear-probing results on four classification datasets.** The Table compares supervised models with DINOv2 pre-trained models in Linear-probing settings on X-ray, CT, and MRI datasets.

| Method | Architecture | NIH Chest X-ray | CheXpert | SARS-CoV-2 | Brain Tumor |
|---|---|---|---|---|---|
| Supervised | DenseNet201 | 0.735 | 0.795 | 0.973 | 0.960 |
| | ResNet152 | 0.718 | 0.779 | 0.936 | 0.948 |
| | VGG19 | 0.696 | 0.750 | 0.891 | 0.933 |
| | ViT-L/16 | 0.751 | **0.829** | **0.983** | 0.975 |
| DINOv2 | ViT-S/14 | 0.747 | 0.805 | 0.943 | 0.962 |
| | ViT-B/14 | 0.755 | 0.812 | 0.922 | 0.972 |
| | ViT-L/14 | **0.763** | 0.821 | 0.950 | 0.974 |
| | ViT-g/14 | 0.759 | 0.818 | 0.978 | **0.976** |

datasets when training only using one to sixteen patients, compared to all patients in the dataset. DINOv2 ViT-L/14 performs worse than an ImageNet-21k pre-trained VIT-L/16 when using less than eight patient instances and slightly outperforms other supervised backbones when eight or more patients are used

On the right of Figure 2, the performance of a frozen DINOv2 model with a U-Net decoder compared to U-Net and TransUnet models is shown when trained on few instances from the Montgomery County dataset. DINOv2 clearly outperforms the other models when using less than eight instances, which is somewhat expected given it was pre-trained while the other models were not.

### 4.3. Parameter-efficient Fine-tuning

We experiment with parameter-efficient fine-tuning (PEFT) techniques on DINOv2 ViT-g/14, which, as a whole, contains 1.1 billion parameters. PEFT methods are used to enable efficient adaptation of large models to downstream tasks, usually achieving performance that is on par with end-to-end fine-tuning while requiring a lot less compute and memory. Previous work by Dutt et al. [44] has highlighted the opportunity of employing PEFT to tune large foundation models for medical image analysis.

We employ two different PEFT techniques: LoRA [37] and BitFit [38]. LoRA is an additive

**Table 5: DINOv2 AUROC performance comparison on large X-ray datasets.** DINOv2 outperforms other methods on Linear-probing and fine-tuning but under performs on kNN evaluations.

| | | NIH Chest X-ray | | | CheXpert | | |
|---|---|---|---|---|---|---|---|
| Method | Architecture | kNN | Linear-probing | Fine-tuning | kNN | Linear-probing | Fine-tuning |
| Supervised | DenseNet201 | **0.675** | 0.735 | **0.769** | 0.783 | 0.795 | 0.882 |
| | Resnet152 | 0.668 | 0.718 | 0.752 | 0.766 | 0.779 | 0.868 |
| | VGG19 | 0.644 | 0.696 | 0.711 | 0.728 | 0.750 | 0.870 |
| | ViT-L/16 | 0.663 | 0.751 | 0.761 | 0.777 | **0.829** | 0.873 |
| MAE | ViT-L/16 | 0.659 | 0.724 | 0.743 | **0.785** | 0.789 | 0.821 |
| CLIP | ViT-L/14 | 0.655 | 0.739 | 0.697 | 0.742 | 0.816 | 0.842 |
| DINOv2 | ViT-L/14 | 0.663 | **0.763** | 0.717 | 0.771 | 0.821 | 0.786 |
| | ViT-g/14 | 0.659 | 0.759 | **0.769** | 0.768 | 0.818 | 0.848 |

**Table 6: Linear vs. U-Net decoder.** Comparison of the Linear and U-Net decoders on the four segmentation tasks used in this analysis. The backbones used in both is a DINOv2 ViT-L/14.

| Decoder | Montgomery County | Shenzhen | AMOS | MSD Heart |
|---|---|---|---|---|
| Linear | 0.945 | 0.930 | 0.515 | 0.703 |
| U-Net | 0.974 | 0.951 | 0.592 | 0.876 |
| $\Delta$ | +2.9 | +2.1 | +7.7 | +17.3 |

method that inserts trainable decomposition matrices in the layers of a transformer, while BitFit is a selective method that unfreezes only the bias terms of the model. Table 8 shows a result and efficiency comparison between the two PEFT methods with a comparison to end-to-end fine-tuning and linear-probing on the NIH Chest X-ray and CheXpert datasets using the DINOv2 ViT-g/14 model.

### 4.4. Cross-task Generalizability Comparison

In this section, we test whether the DINOv2 pre-training results in representations that are much more generalizable across tasks and modalities, compared to other supervised models. To accomplish this task, we carry out two experiments.

First, we compare the segmentation performance of a DINOv2 pre-trained ViT-L/14 with a ViT-L/16 pre-trained with supervised learning on ImageNet21k. Table 9 shows the results. DINOv2

**Table 7: DINOv2 on organ segmentation.** A comparison between using a frozen DINOv2 backbone and other commonly-used segmentation models initialized from scratch.

| Method | Architecture | Montgomery County | Shenzhen | AMOS | MSD Heart |
|---|---|---|---|---|---|
| Scratch | U-Net | 0.973 | **0.951** | 0.432 | **0.911** |
| | TransUnet | **0.974** | **0.951** | 0.535 | 0.892 |
| DINOv2 | ViT-L/14 | **0.974** | **0.951** | **0.592** | 0.876 |
| | ViT-g/14 | 0.973 | 0.950 | 0.540 | 0.88 |

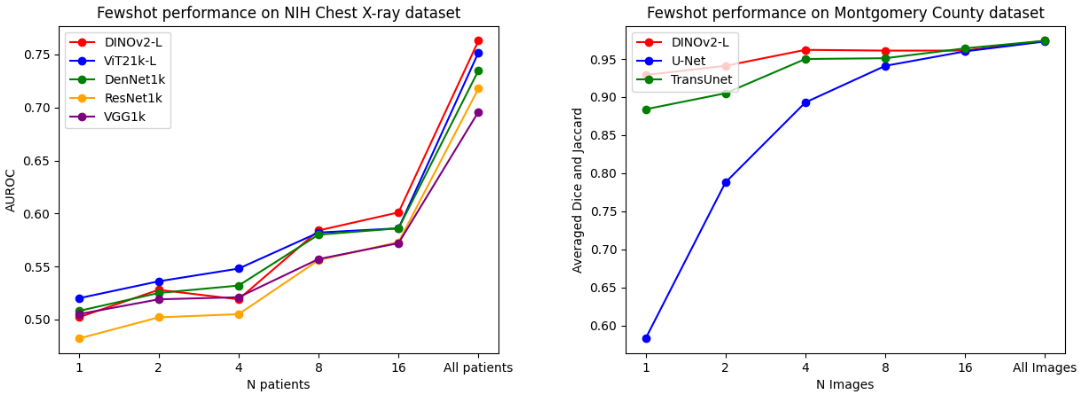


**Figure 2: Few-shot disease classification and organ segmentation.** On the left, DINOv2 is compared with supervised classification models on few-shot disease classification using the NIH Chest X-ray dataset. On the right, a frozen DINOv2 encoder with a U-Net decoder is compared with U-Net and TransUnet models initialized from scratch.

outperforms the supervised ViT on 3 out of the 4 organ segmentation tasks, especially in the challenging AMOS multi-organ segmentation task.

Our second experiment compares the performance of the SAM image encoder with DINOv2 on classification tasks. SAM was trained for prompt-based segmentation and does not have a CLS token. To perform classification with SAM we averaged all the patch embeddings and treated the result as a CLS token. Table 10 shows the results. Images are of size 1024x1024 and only a subset of each dataset was used because of computational limits. DINOv2 significantly outperforms SAM on both datasets, highlighting the cross-task generalizability of the model.

**Table 8: PEFT on DINOv2 ViT-g/14.** Both LoRA and BitFit achieve results that are better than linear-probing while adapting less than 1% of the total parameters.

| Method | Trainable Params. (%) | NIH Chest X-ray | CheXpert |
|---|---|---|---|
| Fine-tuning | 1,100M (100%) | 0.769 | 0.848 |
| Linear-probing | 1,500 (1e-6%) | 0.759 | 0.818 |
| LoRA | 8M (0.7%) | 0.767 | 0.823 |
| BitFit | 0.8M (0.07%) | 0.768 | 0.817 |

**Table 9: DINOv2 vs. ImageNet21k pre-trained ViT on organ segmentation.** DINOv2 outperforms the supervised pre-trained ViT on 3 of the 4 tasks.

| Method | Architecture | Image Size | Montgomery County | Shenzhen | AMOS | MSD Heart |
|---|---|---|---|---|---|---|
| Supervised | ViT-L/16 | 224 | 0.963 | 0.942 | 0.433 | 0.839 |
| DINOv2 | ViT-L/14 | 224 | 0.966 | 0.947 | 0.512 | 0.799 |
|  | ViT-L/14 | 448 | 0.974 | 0.951 | 0.592 | 0.876 |

### 4.5. Qualitative Results

In this section we will show qualitative results of DINOv2 features using principal component analysis (PCA) performed on DINOv2 patch features on X-ray, CT, and MRI scans, following the method delineated in [3]. We will also provide organ segmentation results of linear compared U-Net decoders.

Figure 1 shows the first three PCA components. The PCA is computed between patches of images that are in the same column, and the first 3 components are shown for X-ray, CT, and MRI scans. Thresholding is used on the first PCA component to remove the background. Just like in natural images [3], the colors of the three PCA components correspond well with the same parts of images in the same category. This is an easier task however, compared to natural images, because there is less variability between examinations on medical images compared to natural images.

We also show a visualization of linear and U-Net decoders trained on top of DINOv2 ViT-L/14 features. The linear layer decoder performs surprisingly well, but is limited due to the smaller image size and less adjustable parameters. As expected, the U-Net segmentation results are smoother and represents the ground truth mask more accurately, but is still limited due to the frozen encoder.

**Table 10: SAM vs DINOv2 on X-ray on disease classification.** Only 20,000 samples were used from each dataset and all the images are of size 1024x1024. The average of patch embeddings were used as a CLS token for SAM.

| Method | Architecture | NIH Chest X-ray | CheXpert |
|---|---|---|---|
| SAM | ViT-L/16 | 0.714 | 0.792 |
| DINOv2 | ViT-L/14 | **0.755** | **0.816** |

## 5. Discussion

Foundation models have shown promise for reducing the data annotation problem and increasing model generalizability and robustness. Especially in the medical image analysis domain where data annotation is considerably more expensive than other fields, and where model robustness is critical, these foundation models hold promise towards increasing performance and adoption of deep learning systems in healthcare. The DINOv2 pre-training approach is specifically promising given its ability to learn general-purpose representations and perform well out-of-the-box in linear probing and kNN settings. We believe that using DINOv2 pre-training on medical data is a promising approach for future research, aimed at building large-scale medical foundation models whiteout supervision.

## 6. Reproducibility

All of the data used for this work is publicly available. The code is available using the following link: https://github.com/MohammedSB/DINOv2ForMedical

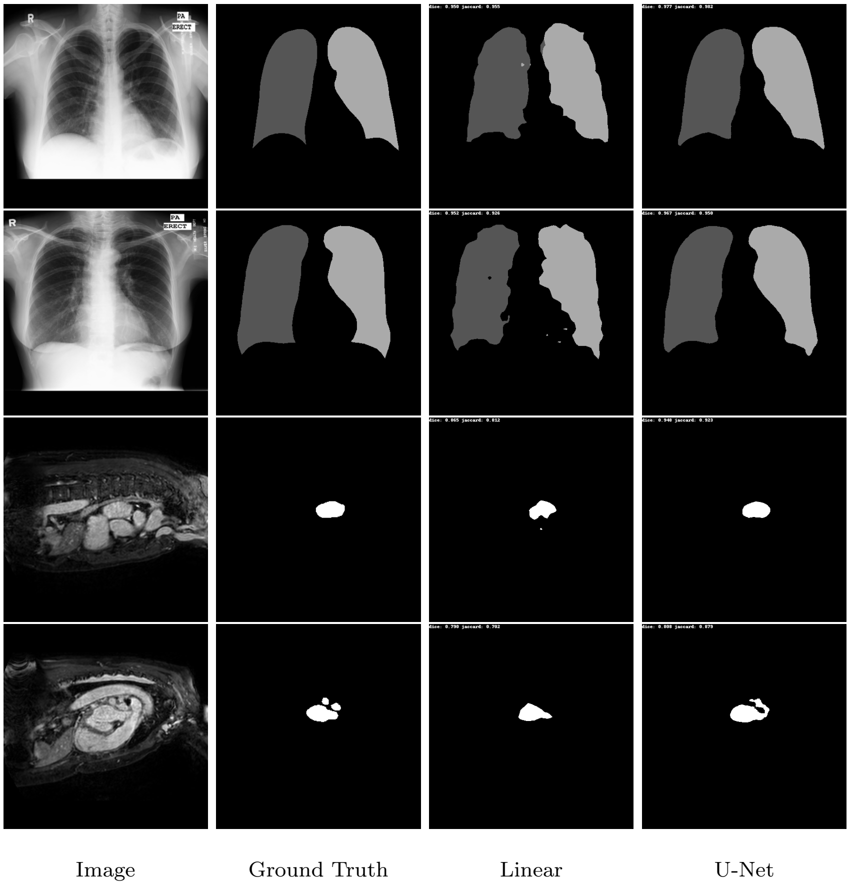
All the logs for training and validation, model weights, and data splits are available using the following link: https://drive.google.com/drive/u/0/folders/1kJpKJIyC-3m3unqm6HmWjhnYGS2jxxwj

## 7. Acknowledgments

This study was conducted as part of the Fatima Fellowship program. We thank the Fatima Fellowship and Hugging Face for organizing and sponsoring the Fatima Fellowship Program.[1]

[1] https://www.fatimafellowship.com/

**Figure 3: Linear vs. U-Net visualization.** The figure shows a qualitative comparison between the linear layer and the U-Net decoder.


R
PA
ERECT
dice: 0.950 jaccard: 0.955
dice: 0.977 jaccard: 0.982
R
PA
ERECT
dice: 0.952 jaccard: 0.926
dice: 0.967 jaccard: 0.950
dice: 0.865 jaccard: 0.812
dice: 0.948 jaccard: 0.923
dice: 0.790 jaccard: 0.702
dice: 0.808 jaccard: 0.879
Image
Ground Truth
Linear
U-Net

We also thank Meta AI for making the DINOv2 and SAM weights and code bases publicly available.

## References

[1] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[3] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[5] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, Denize Atan, Andrew D. Dick, Cathy Williams, Sarah Barman, Jenny H. Barrett, Sarah Mackie, Tasanee Braithwaite, Roxana O. Carare, Sarah Ennis, Jane Gibson, Andrew J. Lotery, Jay Self, Usha Chakravarthy, Ruth E. Hogg, Euan Paterson, Jayne Woodside, Tunde Peto, Gareth Mckay, Bernadette Mcguinness, Paul J. Foster, Konstantinos Balaskas, Anthony P. Khawaja, Nikolas Pontikos, Jugnoo S. Rahi, Gerassimos Lascaratos, Praveen J. Patel, Michelle Chan, Sharon Y. L. Chua, Alexander Day, Parul Desai, Cathy Egan, Marcus Fruttiger, David F. Garway-Heath, Alison Hardcastle, Sir Peng T. Khaw, Tony Moore, Sobha Sivaprasad, Nicholas Strouthidis, Dhanes Thomas, Adnan Tufail, Ananth C. Viswanathan, Bal Dhillon, Tom Macgillivray, Cathie Sudlow, Veronique Vitart, Alexander Doney, Emanuele Trucco, Jeremy A. Guggeinheim, James E. Morgan, Chris J. Hammond, Katie Williams, Pirro Hysi, Simon P. Harding, Yalin Zheng, Robert Luben, Phil Luthert, Zihan Sun, Martin McKibbin, Eoin O'Sullivan, Richard Oram, Mike Weedon, Chris G. Owen, Alicja R. Rudnicka, Naveed Sattar, David Steel, Irene Stratton, Robyn Tapp, Max M. Yates, Axel Petzold, Savita Madhusudhan, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and

Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, September 2023.

[6] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train, 2023.

[7] Josue Ortega Caro, Antonio H. de O. Fonseca, Christopher Averill, Syed A. Rizvi, Matteo Rosati, James L. Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M. Dhodapkar, Chadi G. Abdallah, and David van Dijk. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, 2023.

[8] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, Yuyang Zhao, Xuehui Shi, Junfang Xian, Xiaoxia Qu, Sirui Zhu, Lijie Pan, Xiaoniao Chen, Xiaojia Zhang, Shuai Jiang, Kebing Wang, Chenlong Yang, Mingqiang Chen, Sujie Fan, Jianhua Hu, Aiguo Lv, Hui Miao, Li Guo, Shujun Zhang, Cheng Pei, Xiaojuan Fan, Jianqin Lei, Ting Wei, Junguo Duan, Chun Liu, Xiaobo Xia, Siqi Xiong, Junhong Li, Benny Lo, Yih Chung Tham, Tien Yin Wong, Ningli Wang, and Wu Yuan. Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence, 2023.

[9] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images, 2023.

[10] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023.

[11] Rick Merritt. What are foundation models? https://blogs.nvidia.com/blog/what-are-foundation-models/, March 2023. Accessed: 2023-11-30.

[12] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google usm: Scaling automatic speech recognition beyond 100 languages, 2023.

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[14] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[16] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[17] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

[18] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval, 2020.

[19] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

[20] Yichi Zhang and Rushi Jiao. Towards segment anything model (sam) for medical image segmentation: A survey, 2023.

[21] Pierre Chambon, Christian Bluethgen, Curtis P. Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains, 2022.

[22] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text, 2022.

[23] Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(1):317, December 2019.

[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031, 2019.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[26] Jeya Maria Jose Valanarasu, Yucheng Tang, Dong Yang, Ziyue Xu, Can Zhao, Wenqi Li, Vishal M. Patel, Bennett Landman, Daguang Xu, Yufan He, and Vishwesh Nath. Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training, 2023.

[27] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20730–20740, 2022.

[28] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022.

[29] Bennett Landman, Zhoubing Xu, Juan Eugenio Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *International Workshop on Multimodal Brain Image Analysis*, pages 1–10. Springer, 2015.

[30] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbeláez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Ildoo Kim, Klaus Maier-Hein, Dorit Merhof, Akshay Pai, Beomhee Park,

Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), jul 2022.

[31] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023.

[32] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[33] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[34] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04, 2020.

[35] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

[36] Jun Cheng. brain tumor dataset, 4 2017.

[37] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[38] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly,

Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[40] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[43] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[44] Raman Dutt, Linus Ericsson, Pedro Sanchez, Sotirios A. Tsaftaris, and Timothy Hospedales. Parameter-efficient fine-tuning for medical image analysis: The missed opportunity, 2023.