

# Self-Supervised Pre-Training with Contrastive and Masked Autoencoder Methods for Dealing with Small Datasets in Deep Learning for Medical Imaging

Daniel Wolf<sup>1,2,\*</sup>, Tristan Payer<sup>1</sup>, Catharina Silvia Lisson<sup>2</sup>, Christoph Gerhard Lisson<sup>2</sup>, Meinrad Beer<sup>2</sup>, Michael Götz<sup>2,+</sup>, and Timo Ropinski<sup>1,+</sup>

<sup>1</sup>Visual Computing Research Group, Institute of Media Informatics, Ulm University, Germany

<sup>2</sup>Experimental Radiology Research Group, Department for Diagnostic and Interventional Radiology, Ulm University Medical Center, Germany

<sup>1,2</sup><https://xairad.informatik.uni-ulm.de/>

\*daniel.wolf@uni-ulm.de

+these authors contributed equally to this work

## ABSTRACT

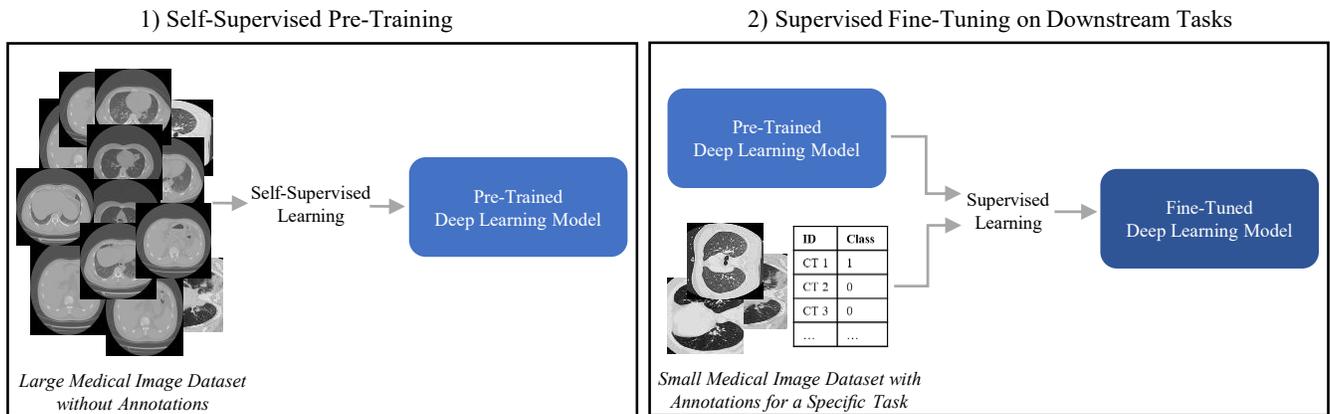
Deep learning in medical imaging has the potential to minimize the risk of diagnostic errors, reduce radiologist workload, and accelerate diagnosis. Training such deep learning models requires large and accurate datasets, with annotations for all training samples. However, in the medical imaging domain, annotated datasets for specific tasks are often small due to the high complexity of annotations, limited access, or the rarity of diseases. To address this challenge, deep learning models can be pre-trained on large image datasets without annotations using methods from the field of self-supervised learning. After pre-training, small annotated datasets are sufficient to fine-tune the models for a specific task. The most popular self-supervised pre-training approaches in medical imaging are based on contrastive learning. However, recent studies in natural image processing indicate a strong potential for masked autoencoder approaches. Our work compares state-of-the-art contrastive learning methods with the recently introduced masked autoencoder approach "Spark" for convolutional neural networks (CNNs) on medical images. Therefore we pre-train on a large unannotated CT image dataset and fine-tune on several CT classification tasks. Due to the challenge of obtaining sufficient annotated training data in medical imaging, it is of particular interest to evaluate how the self-supervised pre-training methods perform when fine-tuning on small datasets. By experimenting with gradually reducing the training dataset size for fine-tuning, we find that the reduction has different effects depending on the type of pre-training chosen. The Spark pre-training method is more robust to the training dataset size than the contrastive methods. Based on our results, we propose the Spark pre-training for medical imaging tasks with only small annotated datasets. We provide ready-to-use code and pre-trained models on GitHub: <https://github.com/Wolfda95/SSL-MedicalImaging-CL-MAE>.

## Introduction

Medical imaging has become an essential part of modern medicine, improving diagnostic and treatment strategies<sup>1</sup>. Deep learning models trained on medical images have recently demonstrated diagnostic accuracy comparable to human experts for narrow clinical tasks<sup>2-5</sup>. This has the potential to reduce the workload of radiologists who are faced with an ever-increasing number of medical images, speed up diagnosis, and minimize the risk of diagnostic errors<sup>6,7</sup>.

The majority of deep learning models are trained by supervised learning, which requires large datasets with annotations for all training samples<sup>8</sup>. However, in medical imaging, annotated datasets for specific tasks are often small due to limited access to the data or the rarity of certain diseases. Further, the complexity of the annotations, which require a high degree of expertise, makes them both costly and time-consuming to obtain<sup>9,10</sup>. In addition to facing challenges with training data, supervised models tend to lack the ability to generalize to different tasks or external institutions, as they mainly learn features correlated with the specific labels rather than learning general feature representations<sup>11</sup>. A promising solution to overcome these challenges is self-supervised learning (SSL)<sup>8</sup>. This is a technique that trains deep learning models to create useful representations from unlabeled datasets. As illustrated in Figure 1, self-supervised learning can be applied in the following way: First, deep learning models are pre-trained with self-supervised learning on a large unlabeled medical image dataset. This allows the model to learn general high-level features of the images. The pre-trained models are then fine-tuned for medical downstream tasks using supervised learning on small annotated datasets. This approach shows remarkable improvements in downstream task performance compared to training the models from scratch<sup>12-17</sup>.

Several self-supervised learning methods have been developed for natural images and used or adapted for medical tasks<sup>8</sup>.



**Figure 1.** Illustration of the self-supervised pre-training and fine-tuning procedure. In the first step, a deep learning model is pre-trained using self-supervised learning on a large unlabeled medical image dataset. In the second step, the pre-trained model is fine-tuned for a medical downstream task using supervised learning on small annotated datasets. The pre-trained model can be fine-tuned for several different downstream tasks.<sup>18–20</sup>

Huang et al.<sup>8</sup> divide them into four categories: “Innate relationship”, “Generative”, “Contrastive” and “Self-prediction”. Their study found that methods in the “Contrastive” category are currently the most popular for medical self-supervised pre-training. In natural image processing, on the other hand, methods from the category “Self-prediction” have gained popularity in recent years<sup>21</sup>. One such method is the masked autoencoder<sup>22,23</sup>. There are two widely used deep learning model categories for imaging tasks: convolutional neural networks (CNNs)<sup>24</sup> and vision transformers (ViTs)<sup>25</sup>. When pre-training vision transformer models, Masked autoencoders have been shown to outperform state-of-the-art contrastive methods<sup>22,23</sup>. However, several publications show that contrastive pre-training remains superior for convolutional models (CNNs)<sup>8,26</sup>.

In a recent study published at the eleventh International Conference on Learning Representations 2023, Tian et al.<sup>26</sup> demonstrate that Masked Autoencoder can be adapted for convolutional models using sparse convolutions. Their new approach, called “SparK”, outperforms all state-of-the-art contrastive methods on a convolutional model, using natural images from ImageNet<sup>27</sup> for self-supervised pre-training.

Today, in the medical imaging domain, convolutional models still remain the most popular models due to their lower computational cost and their robustness to overfitting on smaller datasets<sup>28,29</sup>. Given the popularity of convolutional models in medical imaging, together with the enormous importance of pre-training such models, we believe that the findings of Tian et al.<sup>26</sup> have significant potential for medical imaging. Therefore, in our work, we investigate masked autoencoders for self-supervised pre-training of convolutional models in the medical imaging domain, specifically for CT scans. To the best of our knowledge, we are the first to apply the SparK<sup>26</sup> approach to CT images. We compare state-of-the-art contrastive pre-training methods against SparK by pre-training with a sizeable public CT dataset and performing downstream task evaluations on several CT classification tasks, the benchmark task for evaluating self-supervised pre-training<sup>8</sup>. Furthermore, it is of particular interest how the self-supervised pre-training methods perform on small downstream datasets due to the challenges of obtaining annotated data. Therefore, we gradually reduce the training dataset sizes of our downstream tasks and evaluate the different pre-training methods for each reduction step. We find that the reduction has significantly different effects depending on the type of pre-training. In total, we conducted four pre-trainings, each lasting about 30 days, and over 400 fine-tuning runs on the downstream tasks to evaluate the pre-trainings. Based on our experiments, we conclude with a proposal for self-supervised pre-training on CT images, especially for downstream tasks with small annotated datasets. The code and the pre-trained models are available on GitHub <https://github.com/Wolfda95/SSL-MedicalImaging-CL-MAE>.

## Methods

There are two primary methods for training deep learning models on CT scans: using a 3D model to train on entire CT volumes, or using a 2D model to train on individual slices of the volumes. Each method has its advantages. Training a 3D model on volumes allows the model to better capture the 3D properties of the images<sup>30</sup>. Conversely, training a 2D model on each slice of the volumes separately artificially increases the dataset size, which can improve results for sparse datasets<sup>31–33</sup>. In addition, 2D models require significantly less computation, memory, and time for training, as well as during inference when deployed on hospital machines<sup>30–32,34,35</sup>. While 3D models rely on large GPUs, 2D models can be applied on portable devices after training. Both approaches are widely used in the medical imaging domain, with several recent publications demonstrating

excellent results on both 3D<sup>36,37</sup> and 2D<sup>5,38–40</sup> models for clinically relevant CT imaging tasks. For both types of models, there are several publications investigating self-supervised pre-training with CT images. Tang et al.<sup>14</sup> and Dufumier et al.<sup>16</sup> pre-train 3D models on CT volumes, while Ghesu et al.<sup>12</sup>, Chen et al.<sup>13</sup>, and He et al.<sup>41</sup> pre-train 2D models on CT slices. They all use contrastive methods and achieve significant performance gains on several CT image downstream tasks.

In this work, we have chosen to perform our experiments on 2D models. We see that it is important for deep learning to be accessible worldwide without the need for powerful GPUs. In addition, we believe that the advantages of 2D models for sparse data, discussed earlier, are important, as small annotated datasets remain a critical challenge in medical imaging, even with pre-trained models. However, due to the similar model structure of 2D and 3D convolutional models, and the same pre-training methods that can be applied in the same way in 3D, we assume that the results can be directly transferred to the 3D domain.

In the following, we introduce the self-supervised pre-training methods and the downstream task evaluation procedure. For all our experiments, we choose a ResNet50<sup>42</sup> as our convolutional model due to its widespread use as a baseline for comparisons in vision studies<sup>43</sup> and its popularity in medical image analysis<sup>44</sup>. Our findings are expected to apply to other convolutional models as well.

### Self-Supervised Pre-Training

The first step is to pre-train the model with a sizeable unlabeled image dataset with self-supervised learning. We use the CT slices of the publicly available LIDC-IDRI<sup>18,19</sup> dataset, which contains lung CT scans of 1,010 patients, from which we extract 244,527 CT slices. We use only the CT slices, all other information or labels were excluded. We compare four different self-supervised learning methods, three contrastive methods, and the recent masked autoencoder method SparK<sup>26</sup>. Each of the four pre-trainings runs 900 epochs on an Nvidia GeForce RTX 3090 GPU. Table 1 shows the computational cost of the pre-training.

#### Contrastive Learning

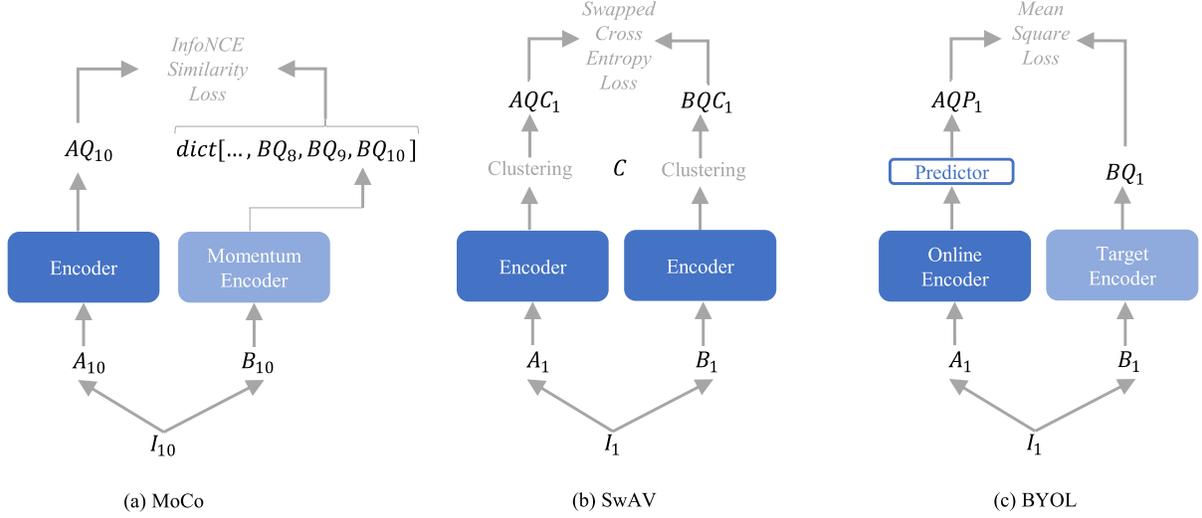
The general idea of Contrastive Learning is as follows: Starting with an unlabeled image dataset, random transforms are applied to the images to obtain several randomly different samples of each original image. The samples are passed through a deep learning model to get latent-space representations of each sample. The encoder model is trained to classify between representations that come from the same original image (positive pairs) and representations that come from different original images (negative pairs).

Following Huang et al.’s<sup>8</sup> study, popular contrastive learning methods from natural image processing that are widely used for medical pre-training are SimCLR<sup>45</sup>, MoCo<sup>46</sup>, SwAV<sup>47</sup>, and BYOL<sup>48</sup>. SimCLR follows the basic contrastive learning strategy, where the model learns to classify between positive and negative pairs within one mini-batch. This requires a large batch size to get enough negative samples in one mini-batch. MoCo, SwAV, and BYOL add additional features to reduce the batch size requirements allowing them to be used with less computational power. They outperform SimCLR on several benchmark tasks<sup>26</sup>. We choose to pre-train our model with MoCo Version 2, SwAV, and BYOL as our baseline contrastive methods. We use the implementations from PyTorch Lightning Bolts<sup>49</sup> with the hyperparameters of the original papers. Details can be found in Supplementary Information. In the following, we explain the three contrastive methods.

**MoCo:** Starting with a dataset of images  $\{I_1, I_2, I_3, \dots\}$ , two random transformations are performed to obtain two randomly different images from each original image:  $\{A_1, A_2, A_3, \dots\}$  and  $\{B_1, B_2, B_3, \dots\}$ . For example, as shown in Figure 2 (a), starting with the original image  $I_{10}$ , the transformed image  $A_{10}$  is computed by an encoder to the latent space representation  $AQ_{10}$ , and the transformed image  $B_{10}$  is computed by a momentum encoder to the latent space representation  $BQ_{10}$ . Both encoders have the same architecture and can be any convolutional deep learning model. The computed latent space representation of the momentum encoder  $BQ_{10}$  is stored in a dictionary together with the latent space representations of the momentum encoder from previous images  $dict[\dots, BQ_7, BQ_8, BQ_9, BQ_{10}]$ . The samples in the dictionary are called keys. The dictionary now has one key that comes from the same original image as the latent space representation of the encoder  $AQ_{10}$ , which is the positive pair and the dictionary has several keys from different original images, which are the negative pairs. The model is trained to classify between positive and negative pairs by applying the InfoNCE<sup>50</sup> loss

$$L_{10} = -\log \frac{\exp(AQ_{10} \cdot BQ_k / \tau)}{\sum_{i=0}^{10} \exp(AQ_{10} \cdot BQ_k / \tau)}, \quad (1)$$

which calculates a similarity score.  $\tau$  is a temperature hyperparameter. Details about the applied transformations and the updating of the weights of the two encoders can be found in<sup>46</sup>. After pre-training, the encoder is used for fine-tuning on downstream tasks. MoCo stores the latent space representations in a dictionary, which can be much larger than a typical mini-batch size. In contrast to SimCLR, where the classification is done between samples in one mini-batch, which must be large enough to get enough negative samples, the dictionary makes MoCo batch size independent. The size of the dictionary is chosen as a hyperparameter.



**Figure 2.** Illustration of the contrastive learning methods MoCo, SwAV, and BYOL for self-supervised pre-training.

MoCo Version 2<sup>51</sup> is an updated version of MoCo that adds an MLP projection head to the encoder and additional data augmentations. The extensions outperform the original model and SimCLR on several benchmark tasks.

**SwAV:** Exactly as in SimCLR, two random transformations are performed on a data set of images  $\underline{I} = \{I_1, I_2, I_3, \dots\}$  in order to obtain two randomly different images from each of the original images:  $\underline{A} = \{A_1, A_2, A_3, \dots\}$  and  $\underline{B} = \{B_1, B_2, B_3, \dots\}$ . Starting with a mini-batch, the transformed images are computed by an encoder, which can be any convolutional model, to a latent space representation:  $\underline{AQ} = \{AQ_1, AQ_2, AQ_3, \dots, AQ_{Bs}\}$  and  $\underline{BQ} = \{BQ_1, BQ_2, BQ_3, \dots, BQ_{Bs}\}$ , with batch size  $Bs$ . The latent space representations are further computed by feature clustering with cluster prototypes  $\underline{C} = \{C_1, C_2, C_3, \dots, C_K\}$ , which leads to the cluster codes

$$\underline{AQC} = \left\{ \frac{1}{\tau} \cdot \underline{AQ}^T \cdot \underline{C}_1, \dots, \frac{1}{\tau} \cdot \underline{AQ}^T \cdot \underline{C}_K \right\} \text{ and } \underline{BQC} = \left\{ \frac{1}{\tau} \cdot \underline{BQ}^T \cdot \underline{C}_1, \dots, \frac{1}{\tau} \cdot \underline{BQ}^T \cdot \underline{C}_K \right\} \quad (2)$$

with the number of prototypes  $K$  as a hyperparameter and the temperature hyperparameter  $\tau$ . The model is trained to predict the cluster codes of transformed images  $\underline{A}$  by the cluster codes of the transformed image  $\underline{B}$  and the other way around by applying a cross-entropy loss with swapped predictions

$$L = - \sum_{k=1}^K \frac{BQC_k}{BQC_k} \cdot \log(\underline{AQC}_k^*) - \sum_{k=1}^K \frac{AQC_k}{AQC_k} \cdot \log(\underline{BQC}_k^*), \quad (3)$$

where the terms  $\underline{AQC}_k^*$  and  $\underline{BQC}_k^*$  are the softmax activation functions applied to the cluster codes. Figure 2 (b) illustrates this procedure for one example image.

The cluster prototypes  $\underline{C}$  are learned during training. The computed cluster codes  $\underline{AQC}$  and  $\underline{BQC}$  of one mini-batch should be equally partitioned by the prototypes. To ensure this equal partitioning and to avoid the trivial solution where all images collapse into the same code, the cluster codes are computed by maximizing the similarity between the latent space representations and the prototypes with the constraint

$$\max_{\underline{AQC}} \text{Tr}(\underline{AQC}^T \underline{C}^T \underline{AQ}) + \varepsilon H(\underline{AQ}), \quad (4)$$

where  $H$  is the entropy and  $\varepsilon$  a regularisation parameter. The same constraint for transform  $B$ .

SwAV further adds a multi-crop strategy to its transforms. The two transformed images  $A$  and  $B$  are obtained by cropping a part of the original image with a larger crop size and 4 additional samples are cropped with a smaller crop size. Details about the multi-crop strategy and further transforms they use can be found in<sup>47</sup>. Similar to SimCLR and in contrast to MoCo, SwAV performs its comparisons within a mini-batch. Thus, SwAV is not completely batch-size independent, however, due to the clustering, the batch-size requirement is lower than in SimCLR, and SwAV outperformed SimCLR on several benchmarks.

**BYOL:** BYOL consists of two encoders, referred to as “online” and “target” networks, that have the same architecture and can be any convolutional model. Again, two random transformations are performed on a dataset of images  $\{I_1, I_2, I_3, \dots\}$  in

order to obtain two randomly different images from each of the original images:  $\{A_1, A_2, A_3, \dots\}$  and  $\{B_1, B_2, B_3, \dots\}$ . As shown in Figure 2 (c), starting with the image pair  $A_1$  and  $B_1$ , image  $A_1$  is computed by the online network to the latent space representation  $AQ_1$  and image  $B_1$  by the target network to  $BQ_1$ . The latent space representation from the online network  $AQ_1$  is further computed by an additional predictor model consisting of an MLP to  $AQP_1$ . The online network, together with the predictor, is trained to predict the latent space representation of the target network  $BQ_1$ , by a mean square loss

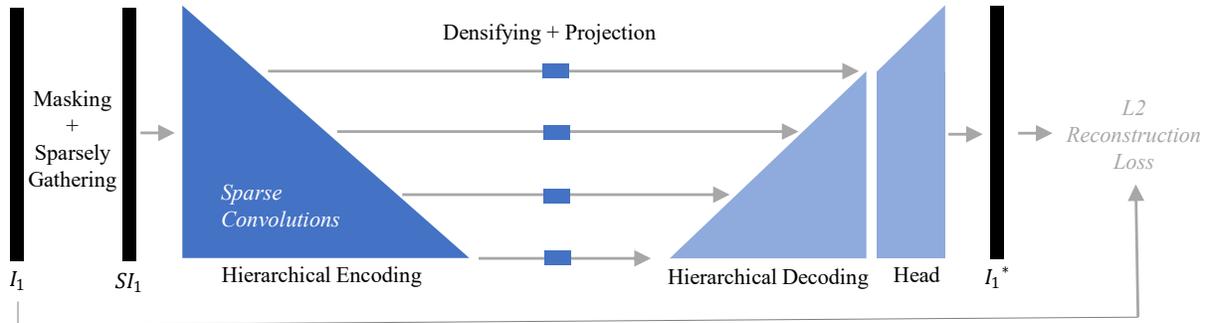
$$L = \|AQP_1 - BQ_1\|_2^2. \quad (5)$$

The target network is updated with a moving average of the parameters from the online network. Since BYOL is only comparing direct pairs, it is batch-size independent.

### Masked Autoencoder

Masked autoencoders are inspired by natural language processing techniques, where models are pre-trained by predicting missing words in a sentence<sup>52</sup>. In the imaging domain, starting from a large unlabeled image dataset, masked autoencoders pre-train deep learning models by dividing the images into patches, masking some of the patches, and training the model to reconstruct the original unmasked images. He et al.<sup>12</sup> show that masked autoencoders outperform state-of-the-art contrastive methods when pre-training vision transformer models. However, applying masked autoencoders to convolutional models (CNNs) showed only moderate success, and contrastive methods remained superior<sup>8,26</sup>. This can be attributed to the characteristics of the models. While transformer models have a variable input size and can drop masked patches, CNNs have a fixed input size and must set masked patches to a fixed value. As evaluated by Tian et al.<sup>26</sup>, the sliding window kernels of CNNs that overlap between masked and non-masked patches result in a loss of mask patterns after several convolutions. They hypothesize that this leads to the moderate success of masked autoencoders for CNNs and try to solve this challenge by using sparse convolutions<sup>53</sup>. This results in a model that skips all masked positions, preventing vanishing mask patterns and ensuring a consistent masking ratio. They use these findings for their self-supervised pre-training approach ‘‘SparK’’. When pre-training a ResNet50<sup>42</sup> model with ImageNet<sup>27</sup> data, SparK outperforms all state-of-the-art contrastive methods<sup>26</sup>. Their approach is the first successful adaption of masked autoencoders to CNNs.

We pre-train our model with CT slices by applying the self-supervised learning method SparK. We use the original PyTorch<sup>54</sup> implementation from the publication. Details can be found in Supplementary Information. In the following, we explain the SparK method:



**Figure 3.** Illustration of the masked autoencoder method SparK for self-supervised pre-training.  $I_1$  is the input image which is divided into patches that are randomly masked and sparsely gathered to the sparse masked image  $SI_1$ . The masked image is computed by a U-Net shape encoder-decoder model that is trained to reconstruct the original image  $I_1$ . The encoder performs sparse convolutions that only compute when the center of a sliding window kernel is covered by a non-masked patch.

**SparK:** Starting with a dataset of images  $\{I_1, I_2, I_3, \dots\}$ , each image is divided into non-overlapping square patches, where each patch is masked independently with a given probability. The probability is a hyperparameter called ‘‘mask ratio’’. The images are converted to sparse images  $\{SI_1, SI_2, SI_3, \dots\}$  by sparsely gathering all unmasked patches. As shown in Figure 3, the SparK model consists of an encoder, which can be any convolutional model and a decoder. The encoder is transformed to perform submanifold sparse convolutions<sup>53</sup>. Submanifold Sparse convolutions only compute when the center of a sliding window kernel is covered by a non-empty element. This causes the encoder to skip all masked places. The decoder is built in a U-Net<sup>55</sup> design with three blocks of upsampling layers, receiving feature maps from the encoder in four different resolutions. This is referred to as ‘‘hierarchical’’ encoding and decoding. The empty parts of the feature maps from the encoder are filled with learnable mask embeddings  $M$  before being computed by the decoder to obtain dense feature maps. This is called ‘‘densifying’’. SparK further adds a projection layer between the encoder and decoder for all computed resolutions in case they have different

**Table 1.** This table shows the computational cost of the pre-training and fine-tuning process. The first row shows the number of trainable parameters in millions. The base is always a ResNet50 encoder with 23.5 million parameters. The pre-training methods BYOL and MoCo consist of two ResNet50 models, BYOL further adds a predictor model, SwAV uses feature clustering with learnable parameters and SparK adds an upsampling decoder to the ResNet50 encoder. For the downstream tasks, only one linear layer is added to the encoder. The second row shows the allocated GPU memory and the third row shows the training time on the GPU (Nvidia GeForce RTX 3090) until the best-performing epoch was reached and the training was terminated.

	Pre-Training				Fine-Tuning on Downstream Task		
	BYOL	SwAV	MoCo V2	SparK	COVID-19	OrgMNIST	Brain
Parameters	70.1 M	24.1 M	47.6 M	25.6 M	23.5 M	23.5 M	23.5 M
GPU Memory	14.4 GB	21,6 GB	18.7 GB	14.4 GB	14.4 GB	14.4 GB	14.4 GB
Training Time	10 d 12 h	11 d 14 h	10 d 7 h	14 d 4 h	6 min	30 min	2 min

network widths. To reconstruct the image, a head module is applied after the decoder with two more upsampling layers to reach the original resolution. The model is trained with an L2 Loss between the predicted images of the model  $\{I_1^*, I_2^*, I_3^*, \dots\}$  and original images  $\{I_1, I_2, I_3, \dots\}$ , computed only on masked positions. After the pre-training, only the encoder is used for the downstream tasks. The sparse convolutions of the encoder can be applied directly to non-masked images without modification since normal convolutions are performed when the images have no masked patches.

### Dataset Preprocessing

The publicly available LIDC-IDRI<sup>18,19</sup> dataset is used for pre-training. The dataset contains lung CT scans of 1,010 patients in DICOM format. In a preprocessing step, each slice of the 3D DICOM images is saved as a PNG file. Therefore, an interval mapping was performed to convert the Hounsfield units (-1024 HU to 3071 HU) of the CT images to grayscale values (0 to 255) for the PNG format. We did not perform intensity windowing because we did not want to focus on a specific area, but instead tried to make the pre-training as general as possible so that it would be applicable to many different downstream tasks. 244,527 slices of the CT scans were extracted and used for pre-training. Intensity normalization was applied to the images using the mean and standard deviation of the dataset.

### Downstream Evaluation

The pre-training performance is evaluated on CT image downstream tasks with the existence of small annotated datasets. The models are initialized with the pre-trained weights and fine-tuned through supervised learning. All pre-trainings ran for 900 epochs. In order to find the best epoch, where the pre-trained weights show the best downstream results, we perform a downstream task evaluation every 50 epochs. The pre-training epoch with the highest F1 score on the downstream tasks is used for all evaluations.

### Downstream Tasks

As demonstrated by Huang et al.<sup>8</sup>, classification tasks serve as a benchmark for evaluating self-supervised pre-training. Typically, only one linear layer is added to the pre-trained encoder to bring the model to the correct output size, resulting in only the weights of one layer not being pre-trained. In contrast, segmentation tasks require adding a large decoder that is added to the pre-trained encoder, such as in a U-Net<sup>55</sup>, resulting in a more significant proportion of untrained model weights. This increases the dependency on the dataset of the downstream task for segmentation tasks. Therefore, to evaluate the performance of the pre-trainings, we focus on classification downstream tasks, although our results are expected to be applicable to other tasks as well.

We selected three classification tasks on CT slices, ensuring that the images do not overlap with those in the pre-training datasets. These tasks include two public challenges and an internal task as part of a clinical study. For all tasks, we have made sure that the slices from the same subject are not used in both training and testing. Our implementations are done in PyTorch Lightning<sup>56</sup> with MONAI<sup>57</sup>, and we trained all tasks on an Nvidia GeForce RTX 3090 GPU using the Adam optimizer with batch-size 64 and learning rate  $10^{-4}$ . We add one linear layer to the pre-trained encoder. Only the linear layer is trained during the first ten epochs before the complete model is fine-tuned. Table 1 shows the computational cost of the fine-tuning. The three tasks are the following:

**COVID-19:** The first task is the public COVID-19 CT Classification Grand Challenge<sup>20</sup>. The provided dataset consists of 349 CT slices from 216 patients with clinical findings of COVID-19 and 397 CT slices from 171 patients without clinical findings

of COVID-19. All slices were selected by senior radiologists at the Tongji Hospital in Wuhan, China. The challenge is a binary classification between COVID-19 findings and no COVID-19 findings. The images of the challenge are already preprocessed and saved in PNG format. The resolutions of the slices range from  $102 \times 137$  to  $1853 \times 1485$ . We resize the slices to  $224 \times 224$  in a preprocessing step in order to obtain the same input size as used for the pre-training. Intensity normalization was applied to the images using the mean and standard deviation of the dataset. 425 slices are used for training, 118 slices for validation, and 203 slices for testing. We perform five downstream training and testing runs with the given data split and report the mean and standard deviation of accuracy, AUC score, and F1 score of the test dataset.

**OrgMNIST:** We use the public OrganSMNIST Challenge from MedMNIST<sup>58</sup> as the second task. The dataset consists of 25,221 image patches cropped around organs from 201 abdominal CT scans. The resolution of the images is  $28 \times 28$ . The challenge is a multi-class classification of 11 body organs. Also for this task, the images are already preprocessed. We resize slices to  $224 \times 224$  and perform intensity normalization. Slices from 115, 16, and 70 CT scans are used as training, validation, and test set, respectively. We perform five downstream training and test runs with the given data split, and the mean and standard deviation of accuracy, AUC score, and F1 score of the test dataset are reported.

**Brain:** The third task is performed on an internal dataset as part of a clinical study about brain hemorrhages. Brain or intracranial hemorrhage is a bleeding inside the skull<sup>59</sup>. It is essential to diagnose these hemorrhages quickly because they can cause various problems for the patient, such as brain infection, brain swelling, or death of brain matter. Hemorrhages occur when blood vessels inside the skull rupture, which can be caused by physical trauma or stroke<sup>59</sup>. For an internal study, CT slices were collected from 100 patients with and 100 patients without a brain hemorrhage. All patients underwent CT examination as part of the routine clinical practice at the University Hospital of Ulm. Representative slices were selected by the two well-trained senior radiologists Dr. Ch. G. Lisson and Dr. Ca. S. Lisson. The aim of this study is to determine whether brain hemorrhages can be detected automatically on CT scans, which could help physicians in their diagnosis. Ethical approval was granted by the Ethics Committee of Ulm University under ID 302/17. More details about the collected slices can be found in Supplementary Information. We participate in this study by using the dataset as a downstream task to classify between brain hemorrhage and no brain hemorrhage. Due to the small size of the dataset, pre-training is essential here. The images are obtained in DICOM format. We performed intensity windowing with a typical brain window with a window width of 80 and a window level of 35. Each slice is resized from  $512 \times 512$  to  $224 \times 224$  and saved in a PNG format. Interval mapping was performed to convert the Hounsfield units to grayscale values for the PNG format and intensity normalization was applied. We perform a five-fold stratified cross-validation using 10% of the training data for validation and report the mean and standard deviation of accuracy, AUC score, and F1 score over the five runs.

### **Downstream Dataset Reduction**

As discussed earlier, annotated datasets for specific tasks in medical imaging are often small due to the high complexity of annotations, the limited access to data, or the rarity of diseases. Pre-training is crucial for such small datasets. We want to evaluate if the pre-training methods perform differently depending on the downstream training dataset size. We attempt to find the pre-training method that is best suited for small downstream datasets and evaluate how many samples per class are needed to achieve decent downstream results.

Our three downstream tasks have different dataset sizes. While we use only 145 images as the training set of the binary classification task Brain, the training dataset of the binary classification task COVID-19 consists of 425 images, and the training dataset of the eleven classes classification task OrgMNIST consists of 13,952 images, resulting in approximately 1,268 images per class. We randomly reduce the training datasets of each downstream task in steps of 25%. For each reduction step, we compare the four self-supervised pre-training methods, BYOL, MoCoV2, SwaV, and SparK, by fine-tuning the pre-trained models with the reduced training datasets. We perform reduction steps until the training datasets are too small to achieve decent fine-tuning results. We choose an F1 score of 0.7 on the test datasets as the limit since a lower F1 score leads to an amount of variance and randomness that are too high to make a proper comparison. This means if the F1 score of a downstream task is below 0.7 for all pre-training methods, we do not perform a further reduction step for this downstream task. For all reduction steps, we calculate the mean over five fine-tuning runs for each downstream task with each pre-training method, as described in chapter [Downstream Tasks](#). For each of the five runs, the random reduction is made with different seeds to get different remaining samples for each run, ensuring a similar distribution of classes as in 100% of the data.

## **Results**

### **Downstream Results**

Table 2 shows the results for the three downstream tasks, COVID-19, OrgMNIST, and Brain when fine-tuning on the complete training datasets. For all tasks, we first list the results of training the ResNet50 model with a random initialization from PyTorch

**Table 2.** This table compares the self-supervised pre-training methods BYOL, SwAV, MoCoV2, and SparK, by fine-tuning a ResNet50 on the three downstream tasks COVID-19, OrgMNIST and Brain. The AUC and F1 scores are mean and standard deviations over five fine-tuning runs. The first line shows the baseline results of training the ResNet50 from scratch without pre-training.

Pre-Train	Downstream COVID-19		Downstream OrgMNIST		Downstream Brain	
	AUC Score	F1 Score	AUC Score	F1 Score	AUC Score	F1 Score
-	0.737 ± 0.033	0.679 ± 0.033	0.971 ± 0.001	0.755 ± 0.003	0.678 ± 0.037	0.447 ± 0.157
BYOL	<b>0.854 ± 0.001</b>	0.767 ± 0.002	0.979 ± 0.005	0.790 ± 0.005	0.734 ± 0.142	0.606 ± 0.062
SwAV	0.807 ± 0.006	0.744 ± 0.013	0.972 ± 0.003	0.769 ± 0.003	0.734 ± 0.046	0.609 ± 0.072
MoCoV2	0.824 ± 0.005	<b>0.780 ± 0.009</b>	<b>0.981 ± 0.001</b>	<b>0.817 ± 0.001</b>	0.825 ± 0.010	0.770 ± 0.064
SparK	0.828 ± 0.006	0.776 ± 0.009	<b>0.981 ± 0.001</b>	0.808 ± 0.003	<b>0.919 ± 0.015</b>	<b>0.812 ± 0.080</b>

without pre-training, followed by the results of using self-supervised pre-training on the LIDC-IDRI dataset with the contrastive methods BYOL, MoCoV2, and SwAV and the new masked autoencoder method SparK.

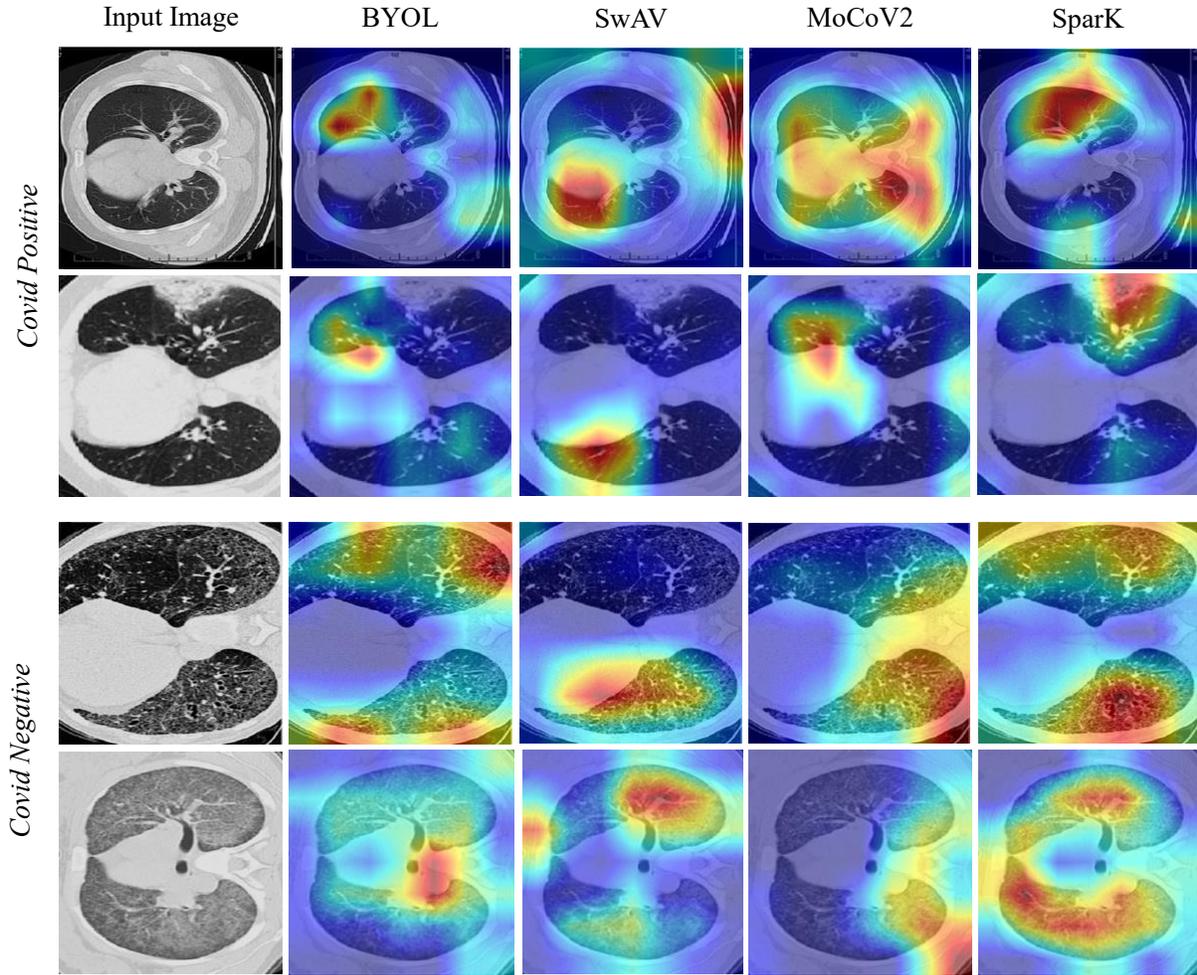
All pre-training methods yield significantly better results than training the model with a random initialization without pre-training. For the Brain task with a relatively small downstream dataset, the SparK method outperforms all contrastive methods with large improvements. For instance, the AUC score improves by 9.4% compared to MoCoV2 and 18.5% compared to SwAV. However, for the COVID-19, and OrgMNIST tasks with larger downstream datasets, the results of the different pre-training methods are more closely aligned. BYOL achieves the highest AUC score on the COVID-19 task, outperforming SparK by 2%, while MoCoV2 has the highest F1 score, outperforming SparK by 0.4%. For the OrgMNIST task SparK and MoCoV2 both reach the same level of AUC score.

We further applied Grad-Cam<sup>60,61</sup>, a technique for generating visual explanations for the decisions of CNNs, to the test dataset images of the COVID-19 task. Figure 4 shows two example images for both classes, covid positive and covid negative. The first column shows the input images, followed by attention heatmaps that highlight the important regions in the input image for the decision of the model. We compare the model’s attention of the four pre-training methods BYOL, SwAV, MoCoV2, and SparK after fine-tuning. A qualitative analysis of the Grad-Cam generated heatmaps reveals large differences between the main focus points of the models depending on the chosen pre-training method. We evaluated the differences between the pre-training methods by calculating the correlation between two heatmaps of the same input image from different pre-training methods. The mean correlations between the different methods are between 0.02 and 0.07, which shows a large variation of the focus points. However, the majority of focus points are located in the lung, as expected.

### Downstream Dataset Reduction

Table 3 shows the sizes of the training datasets after each reduction step for the three downstream tasks COVID-19, OrgMNIST, Brain, until the limit of an F1 score below 0.7 is reached for all pre-training methods. In the brackets are the approximate number of images per class. The exact number cannot be provided since we perform a random reduction for each of the five fine-tuning runs, which only ensures a similar distribution of classes as in 100% of the data. For the Brain task, with the smallest dataset, we reach the limit of an F1 score below 0.7 after a reduction to 75% of the training data, which is approximately 54 images per class. For the COVID-19 task, we perform reductions to 75%, 50%, and 25% until we reach the limit. For the OrgMNIST task, with the largest training dataset, we are still above the limit by reducing the dataset to 25%. We perform further reductions to 10% and 5% of the training data. For all three tasks, we reach the 0.7 F1 score limit, with approximately 50 to 70 samples per class, as shown in Table 3.

Figure 5 shows the downstream task results for the reduction steps listed in table 3, comparing the four pre-training methods BYOL, SwAV, MoCoV2, and SparK. The results indicate that reducing the downstream training data set has different effects depending on the type of pre-training chosen. For the COVID-19 task, the AUC and F1 score of the SparK pre-training method remain constant up to a reduction of 50% (approximately 106 samples per class). Meanwhile, the performance of BYOL and MoCoV2, which yield similar or better results than SparK for 100% of the training data, decreases with the training dataset reduction. After a reduction to 50% of the training dataset, SparK outperforms all other pre-training methods. On the OrgMNIST task, with the largest training dataset, the SparK’s AUC score remains constant until a reduction to 25% (approximately 206 samples per class) and a further reduction to 10% of the training data (approximately 126 samples per



**Figure 4.** Example images of heatmaps generated with Grad-Cam when passing the input image<sup>20</sup> through a fine-tuned ResNet50 model that was pre-trained with the four different methods BYOL, SwAV, MoCoV2, and SparK.

**Table 3.** This table shows the size of the training datasets for each reduction step of the three downstream tasks COVID-19, OrgMNIST, and Brain. For each step, a ResNet50 is fine-tuned, initialized with the four different pre-training methods BYOL, SwAV, MoCoV2, and SparK, and evaluated on the test dataset. If, after a reduction step, the F1 score is below 0.7 with all pre-training methods, no further reduction is performed for this downstream task. The approximate number of images per class for each reduction step is in the brackets.

Downstream COVID-19		Downstream OrgMNIST		Downstream Brain	
Portion	Size	Portion	Size	Portion	Size
100%	425 (≈ 212 per class)	100%	13,952 (≈ 1,268 per class)	100%	145 (≈ 72 per class)
75%	318 (≈ 190 per class)	75%	10,464 (≈ 951 per class)	75%	108 (≈ 54 per class)
50%	212 (≈ 106 per class)	50%	6976 (≈ 634 per class)		
25%	106 (≈ 53 per class)	25%	2488 (≈ 226 per class)		
		10%	1395 (≈ 126 per class)		
		5%	697 (≈ 63 per class)		

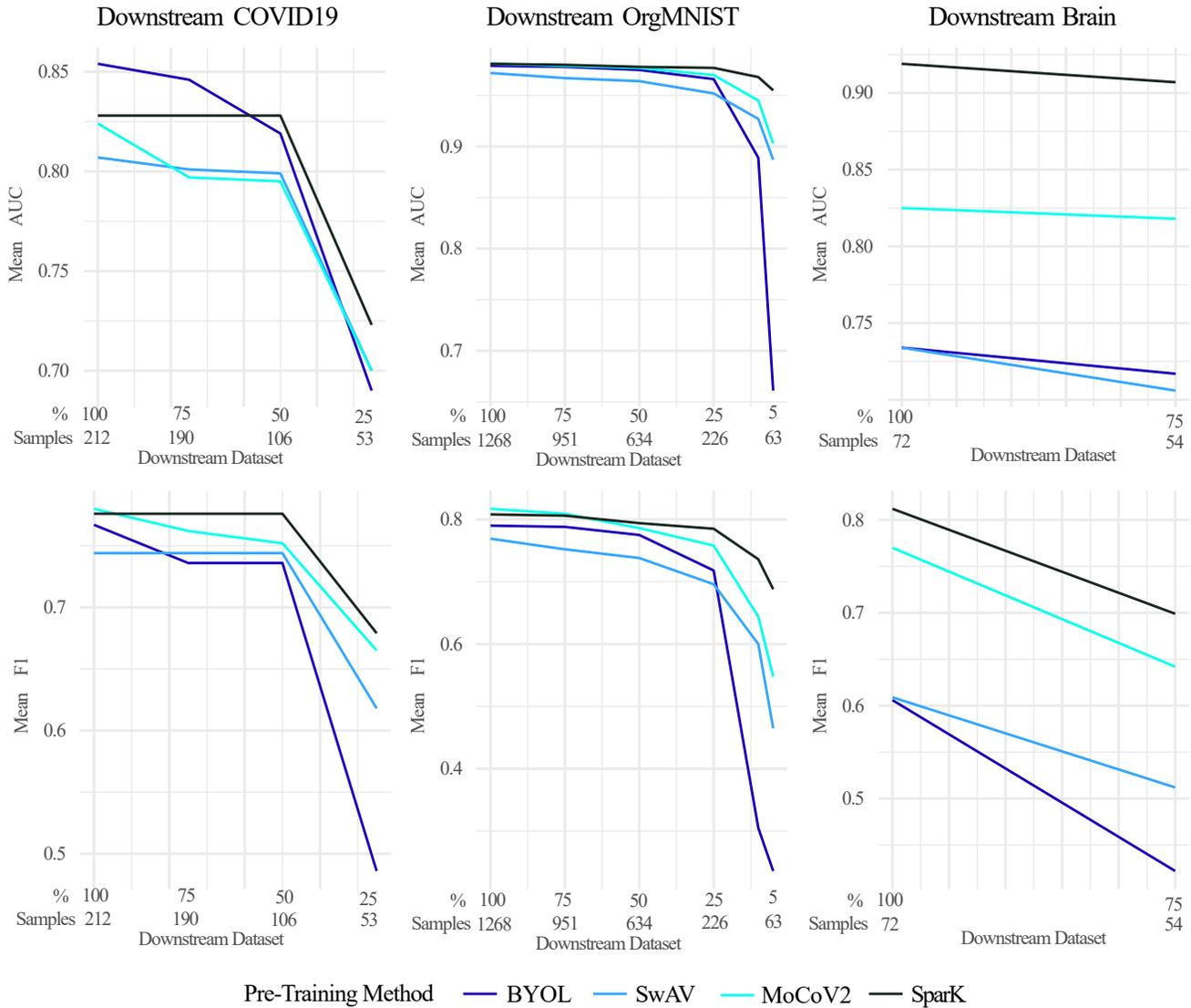
class) results in a performance loss of only 0.015 AUC score. All other methods show larger performance losses in AUC score when reduced to 10% of the training data: BYOL 9%, SwAV 4.5%, and MoCoV2 3.6%. SparK outperforms all other methods after a reduction to 50% of the training data. For both tasks, COVID-19 and OrgMNIST, BYOL has the largest performance loss when reducing the training datasets. The Brain task has the smallest training dataset. The reduction to 75% does not show significant differences depending on the pre-training method. SparK outperforms all other methods with 100% of the training data and is still the best-performing method after the reduction.

## Discussion

Our particular focus in this work has been to find solutions to the challenge of having only small annotated datasets in the medical imaging domain, specifically on CT scans. Convolutional neural networks (CNNs) have proven themselves to be more robust to overfitting on smaller datasets compared to vision transformers, which is one reason why they are still the most popular models in the medical imaging domain today<sup>28,29</sup>. Furthermore, a widely used approach for deep learning on CT scans is to train 2D models by using each slice of the CT scans separately<sup>5,38-40</sup>, which can improve the performance on small datasets, due to the artificially increased dataset size<sup>9,10</sup>. In addition to these two points, an essential part of dealing with small annotated datasets is a well-suited self-supervised pre-training with large unannotated CT image datasets. While the self-supervised pre-training method masked autoencoder is very successful on vision transformers, contrastive learning remained the best-performing method on CNNs due to the model properties<sup>8,26</sup>. SparK<sup>26</sup> is the first method to successfully adapt masked autoencoders to CNNs, making them compatible with contrastive methods. SparK's performance is demonstrated on natural images from ImageNet<sup>27</sup>. Many methods from natural images can be directly transferred to the medical domain. However, not all methods show the same behavior since medical images have different structures and color schemes<sup>63</sup>. Therefore in our work, we evaluated SparK for the medical imaging domain on CT slices and compared it to state-of-the-art contrastive methods. We selected three downstream tasks that include CT scans from different hospitals, scanners, and body parts to prove the generalizability of the self-supervised pre-training.

Fine-tuning on 100% of the downstream training datasets did not reveal a single pre-training method that performs best on all tasks. On the COVID-19 task, BYOL and MoCoV2 show the best performance, while MoCoV2 and Spark show the best performance on the OrgMNIST task, and Spark clearly outperforms all other methods on the Brain task. To continue our attempt to find solutions to the challenge of having only small annotated datasets, we gradually reduced the training dataset sizes of our downstream tasks. The results on the COVID-19 and the OrgMNIST tasks show that SparK is more invariant to the training dataset size compared to the contrastive methods, where we see more considerable performance losses with reduced training data. When reduced to approximately less than 100 to 150 images per class, SparK outperforms all other methods on both tasks. This also explains why Spark performs best on the Brain task for 100% of the training data. Brain has the smallest training dataset, with only about 72 samples per class. Based on these results, we propose self-supervised learning with SparK for CT image classification tasks with less than 150 samples per class. Furthermore, we suggest having at least 60 images per class to achieve decent results with an F1 score above 0.7. However, the minimum number of samples needed to achieve decent results, as well as the number of samples needed for SparK to outperform the other pre-training methods, is highly dependent on the task and the dataset and cannot be generalized. Thus, these numbers give only a rough indication. In general, our results show that the smaller the number of samples, the better SparK performs compared to other methods. To the best of our knowledge, we are the first to discover different downstream performances with dataset reduction depending on the pre-training method in medical imaging. In the future, it would be interesting to investigate these findings further to understand better why SparK is more dataset-reduction invariant. In general, masked autoencoders like SparK focus more on learning local relations in the images to perform the reconstruction task, while contrastive learning focuses more on the relationship between different images. One explanation for SparK's larger dataset reduction invariance could be that internal relations are more relevant for medical images, especially when the downstream dataset is small, and thus, the pre-training is more important because there is less data to change the pre-training weights of the model with supervised fine-tuning.

A limitation of our work is that we performed our evaluation only on one pre-training dataset. However, works from natural image processing show that findings on one pre-training dataset are transferable to other datasets in the same domain<sup>21</sup>. The evaluation is mainly done on ImageNet<sup>27</sup> and later transferred to other natural image datasets<sup>21</sup>. After our work has shown the great potential of SparK for CT imaging, especially for small annotated datasets, it would be interesting to evaluate SparK on further pre-training datasets and other downstream tasks such as segmentation or object detection. Furthermore, an investigation of pre-training and fine-tuning on MRI images could be promising.



**Figure 5.** This figure compares the performance of the self-supervised pre-training methods BYOL, SwAV, MoCoV2, and SparK for small downstream datasets. The pre-trained ResNet50 models are fine-tuned on the three downstream tasks COVID-19, OrgMNIST and Brain with gradually reducing the downstream training datasets until the datasets are too small to achieve decent fine-tuning results. The top row shows the evolution of the mean AUC score, and the bottom row shows the evolution of the mean F1 score. The x-axis provides the percentage and the approximate number of samples per class of the training datasets.<sup>62</sup>

## Conclusion

Self-supervised pre-training is an essential tool in medical imaging to be able to train deep learning models on only small annotated datasets. In our work, we evaluated different self-supervised pre-training methods for CT imaging tasks. We compared state-of-the-art contrastive learning methods, which are currently the most popular self-supervised pre-training methods in medical imaging, with the recently introduced masked autoencoder method SparK. Our focus was to find the best-suited method, especially for downstream tasks with small annotated datasets. Our results show that the SparK pre-training method is more invariant to the downstream training dataset size compared to the contrastive methods, where we see more considerable performance losses as the downstream training datasets become smaller. Based on these findings, we propose the SparK pre-training method for CT imaging tasks with only small annotated datasets. We believe that SparK has great potential in the medical imaging domain since small annotated datasets are a common challenge when dealing with medical data. Pre-trained deep learning models with the SparK method can be used for many medical tasks to obtain models that can assist radiologists in minimizing the risk of diagnostic errors, reducing the radiologist's workload, or speeding up diagnosis.

## References

1. Hong, A. S. *et al.* Trends in diagnostic imaging utilization among medicare and commercially insured adults from 2003 through 2016. *Radiology* **294**, 342–350 (2020).
2. Dunnmon, J. A. *et al.* Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology* **290**, 537–544 (2019).
3. Park, A. *et al.* Deep learning–assisted diagnosis of cerebral aneurysms using the headxnet model. *JAMA network open* **2**, e195600–e195600 (2019).
4. Bien, N. *et al.* Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of mrnet. *PLoS medicine* **15**, e1002699 (2018).
5. Wang, X. *et al.* A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans. *NeuroImage: Clin.* **32**, 102785 (2021).
6. Lantsman, C. D. *et al.* Trend in radiologist workload compared to number of admissions in the emergency department. *Eur. J. Radiol.* **149**, 110195 (2022).
7. Alonso-Martínez, J. L., Sánchez, F. A. & Echezarreta, M. U. Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. *Eur. journal internal medicine* **21**, 278–282 (2010).
8. Huang, S.-C. *et al.* Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit. Medicine* **6**, 74 (2023).
9. Maier-Hein, L. *et al.* Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. communications* **9**, 5217 (2018).
10. Kiryati, N. & Landau, Y. Dataset growth in medical image analysis research. *J. imaging* **7**, 155 (2021).
11. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15**, e1002683 (2018).
12. Ghesu, F. C. *et al.* Contrastive self-supervised learning from 100 million medical images with optional supervision. *J. Med. Imaging* **9**, 064503 (2022).
13. Chen, X., Yao, L., Zhou, T., Dong, J. & Zhang, Y. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern recognition* **113**, 107826 (2021).
14. Tang, Y. *et al.* 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*, 20730–20740 (2022).
15. Truong, T., Mohammadi, S. & Lenga, M. How transferable are self-supervised features in medical image classification tasks? In *Machine Learning for Health*, 54–74 (PMLR, 2021).
16. Dufumier, B. *et al.* Contrastive learning with continuous proxy meta-data for 3d mri classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 58–68 (Springer, 2021).
17. Ewen, N. & Khan, N. Targeted self supervision for classification on a small covid-19 ct scan dataset. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1481–1485 (IEEE, 2021).

18. Armato III, S. G. *et al.* The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Med. physics* **38**, 915–931 (2011).
19. Armato III, S. G. *et al.* Data from lidc-idri [data set]. *The Cancer Imaging Arch.* DOI: [10.7937/K9/TCIA.2015.LO9QL9SX](https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX) (2015).
20. Yang, X. *et al.* Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865* (2020).
21. Balestrieri, R. *et al.* A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210* (2023).
22. He, K. *et al.* Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (2022).
23. Xie, Z. *et al.* Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9653–9663 (2022).
24. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Adv. neural information processing systems* **25** (2012).
25. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
26. Tian, K. *et al.* Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations* (2023).
27. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. journal computer vision* **115**, 211–252 (2015).
28. Kshatri, S. S. & Singh, D. Convolutional neural network in medical image analysis: a review. *Arch. Comput. Methods Eng.* **30**, 2793–2810 (2023).
29. Suganyadevi, S., Seethalakshmi, V. & Balasamy, K. A review on deep learning in medical image analysis. *Int. J. Multimed. Inf. Retr.* **11**, 19–38 (2022).
30. Avesta, A. *et al.* Comparing 3d, 2.5 d, and 2d approaches to brain image auto-segmentation. *Bioengineering* **10**, 181 (2023).
31. Zettler, N. & Mastmeyer, A. Comparison of 2d vs. 3d u-net organ segmentation in abdominal 3d ct images. In *International Conference on Computer Graphics, Visualization and Computer Vision 2021 - WSCG* (2021).
32. Kern, D., Klauck, U., Ropinski, T. & Mastmeyer, A. 2d vs. 3d u-net abdominal organ segmentation in ct data using organ bounds. In *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11601, 192–200 (SPIE, 2021).
33. Bhattacharjee, R. *et al.* Comparison of 2d and 3d u-net breast lesion segmentations on dce-mri. In *Medical Imaging 2021: Computer-Aided Diagnosis*, vol. 11597, 81–87 (SPIE, 2021).
34. Yu, J. *et al.* 2d cnn versus 3d cnn for false-positive reduction in lung cancer screening. *J. Med. Imaging* **7**, 051202–051202 (2020).
35. Nemoto, T. *et al.* Efficacy evaluation of 2d, 3d u-net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi. *J. radiation research* **61**, 257–264 (2020).
36. Lisson, C. S. *et al.* Deep neural networks and machine learning radiomics modelling for prediction of relapse in mantle cell lymphoma. *Cancers* **14**, 2008 (2022).
37. Andrearczyk, V. *et al.* Overview of the hecktor challenge at miccai 2020: automatic head and neck tumor segmentation in pet/ct. In *Head and Neck Tumor Segmentation: First Challenge, HECKTOR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings I*, 1–21 (Springer, 2021).
38. Jiang, M. *et al.* Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 196–206 (Springer, 2022).
39. Xing, X. *et al.* Cs 2: A controllable and simultaneous synthesizer of images and annotations with minimal human intervention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 3–12 (Springer, 2022).
40. Baghdadi, N. A. *et al.* An automated diagnosis and classification of covid-19 from chest ct images using a transfer learning-based convolutional neural network. *Comput. biology medicine* **144**, 105383 (2022).
41. He, X. *et al.* Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medrxiv* 2020–04 (2020).

42. 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*, 770–778 (2016).
43. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
44. Kora, P. *et al.* Transfer learning techniques for medical image analysis: A review. *Biocybern. Biomed. Eng.* **42**, 79–107 (2022).
45. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
46. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738 (2020).
47. Caron, M. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. *Adv. neural information processing systems* **33**, 9912–9924 (2020).
48. Grill, J.-B. *et al.* Bootstrap your own latent—a new approach to self-supervised learning. *Adv. neural information processing systems* **33**, 21271–21284 (2020).
49. Borovec, J., Falcon, W., Nitta, A. *et al.* Lightning-ai/lightning-bolts: 0.5.0 release, DOI: [10.5281/zenodo.7447212](https://doi.org/10.5281/zenodo.7447212) (2022).
50. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
51. Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).
52. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
53. Graham, B. & Van der Maaten, L. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307* (2017).
54. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32** (2019).
55. 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*, 234–241 (Springer, 2015).
56. Falcon, W., Borovec, J. *et al.* Pytorchlightning/pytorch-lightning: 0.7.6 release, DOI: [10.5281/zenodo.3828935](https://doi.org/10.5281/zenodo.3828935) (2020).
57. Consortium, M. Monai: Medical open network for ai: 1.0.0 release, DOI: [10.5281/zenodo.7086266](https://doi.org/10.5281/zenodo.7086266) (2022).
58. Yang, J. *et al.* Medmnist v2—a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Sci. Data* **10**, 41 (2023).
59. Qureshi, A. I. *et al.* Spontaneous intracerebral hemorrhage. *New Engl. J. Medicine* **344**, 1450–1460 (2001).
60. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).
61. Gildenblat, J. & contributors. Pytorch library for cam methods. <https://github.com/jacobgil/pytorch-grad-cam> (2021).
62. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021).
63. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. *Adv. neural information processing systems* **32** (2019).

## Acknowledgements

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

## Author contributions statement

D.W. conceived the experiments, D.W. and T.P. conducted the experiments, C.S.L and C.G.L. collected and prepared a medical dataset, D.W. and M.B. and T.R. and M.G. analyzed the results. All authors reviewed the manuscript.

## **Funding**

This work is funded by “NUM 2.0” (FKZ: 01KX2121) as part of the Racoon Project.

## **Data availability statement**

Pre-Training: The LIDC-IDRI<sup>18,19</sup> dataset is available for public use under the license CC BY 3.0.

Downstream: The COVID-19 CT Classification Grand Challenge<sup>20</sup> dataset is available at <https://github.com/UCSD-AI4H/COVID-CT>; The OrganSMNIST dataset from MedMNIST<sup>58</sup> is available for public use under the license CC BY 4.0; The internal Brain dataset cannot be made publicly available due to strict data security restrictions.

## **Ethics declarations**

For the internal Brain task, ethical approval was granted by the Ethics Committee of Ulm University under ID 302/17.

## **Competing interests**

The authors declare no competing interests.

## Supplementary Information

### Pre-Training Hyperparameter

We use the hyperparameter of the original papers. Only the batch size is adapted due to GPU constraints. We use the maximum possible batch size on our GPU.

**Table 4.** Hyperparameter MoCo

Parameters	Values
Input Size	$512 \times 512$
Transforms	crop, horizontal flip, gaussian blur
Number of Crops	2
Size of Crops	$224 \times 224$
Optimizer	SGD
Batch Size	64
Learning Rate	$1e-4$
Momentum	0.9

**Table 5.** Hyperparameter SwAV

Parameters	Values
Input Size	$512 \times 512$
Transforms	gaussian blur, crop
Number of Crops	2; 6
Size of Crops	$224 \times 224$ ; $96 \times 96$
Min Scale Crops	0.90; 0.10
Max Scale Crops	1.0; 0.33
Optimizer	Lars
Batch Size	128
Learning Rate	0.15
Weight Decay	$1e-6$
Sinkhorn Iterations	3
Number Cluster Prototypes	500
Freeze Cluster Prototypes	313

**Table 6.** Hyperparameter BYOL

Parameters	Values
Input Size	$512 \times 512$
Transforms	crop, horizontal flip, gaussian blur
Number of Crops	2
Size of Crops	$224 \times 224$
Optimizer	LARS
Batch Size	64
Learning Rate	$1e-3$
Weight Decay	$1.5e-6$

**Table 7.** Hyperparameter SparK

Parameters	Values
Input Size	$512 \times 512$
Patch Size	$32 \times 32$
Mask Ratio	60%
Augmentations	horizontal flip, crop
Batch Size	32
Optimizer	LAMB
Learning rate	Cosine Annealing (peak: $25e-6$ )

### 0.1 CT Images of the Brain Downstream Task

**Table 8.** Downstream Task Brain:

Parameters	Values
Format	DICOM
Size	$512 \times 512$
Slice Thickness	1 mm
Area	Brain
Window Center	35/700 HU
Window Width	80/3020 HU
Tube voltage	100-120 kV
CTDI	33-45
DLP	490-805 mgy·cm
Type	No Contrast-Enhanced
Kernel	Soft Tissue
Scanners	PHILIPS Brilliance iCT 256 Siemens Somatom Definition AS+ Siemens Somatom Edge Plus
Gender	Unknown (anonymization)
Age	Unknown (anonymization)