How Far Does Contrastive Self-Supervision Get You on Breast Ultrasound? A Label-Efficiency Study on BUSI

Asma Zizaan[1][0000−0002−9591−6550] and Ali Idri[2][0000−0002−4586−4158]

[1] Faculty of medical sciences, UM6P Hospitals, Mohammed VI Polytechnic university, Benguerir, Morocco
[2] ENSIAS, Mohammed V University, Rabat, Morocco
asma.zizaan@um6p.ma ali.idri@ensias.um5.ac.ma

Abstract. Breast ultrasound is a low-cost, widely accessible screening modality. However, the development of accurate deep learning classifiers for breast ultrasound is limited by the lack of expert-annotated data. Self-supervised learning (SSL) offers a pathway to high-quality representations without labels, and is widely reported to be most useful precisely when labeled data is scarce. We validate this premise on the public BUSI breast ultrasound dataset using SimCLR, assessing both linear probing and full fine-tuning on a labeled-data efficiency curve (10%–100% of available labels) against two baselines: an encoder trained from scratch (random initialization) and an ImageNet-supervised encoder (fully fine-tuned). Contrary to the common low-label narrative, we find that a randomly initialized ResNet-50 trained on as few as 62 images already reaches a macro AUC of 0.863, and that ImageNet-supervised fine-tuning outperforms SimCLR at every label fraction tested, including the lowest (10%: 0.962 vs. 0.747 AUC). SimCLR fine-tuning still consistently outperforms linear probing and yields non-trivial representations from zero labels. Increasing the label budget reduces the gap to supervised pretraining, but does not close it. We argue this is a property specific to BUSI and SimCLR in this configuration: findings may not generalise to other ultrasound datasets or SSL paradigms. We complement these quantitative results with Grad-CAM analysis of attention patterns across methods, and release all code, configurations, and pretrained checkpoints to enable reproducibility.

Keywords: Self-supervised learning · Breast ultrasound · Contrastive learning · Label efficiency · Explainable AI

1 Introduction

Breast cancer remains the most commonly diagnosed cancer among women worldwide, and early detection through screening is central to improving survival outcomes. Ultrasound is particularly valuable in this context: it is radiation-free, low-cost, and widely deployed even in resource-constrained settings, making it

a natural target for AI-assisted screening tools. However, the development of robust deep learning classifiers for breast ultrasound is fundamentally limited by data scarcity , public benchmark datasets such as BUSI [1] contain only a few hundred annotated images, and expert annotation of medical images is costly and requires scarce clinical expertise.

Self-supervised learning (SSL) has emerged as a promising strategy to address this bottleneck by learning transferable representations directly from unlabeled data. Contrastive learning, exemplified by SimCLR [2], learns representations by maximizing agreement between augmented views of the same image, and has shown strong results on natural image benchmarks such as ImageNet. A widely repeated claim in the medical imaging literature is that SSL pretraining is most valuable precisely when labeled data is scarce [11,12] , intuitively appealing, since clinical annotation is the binding constraint in practice. However, this claim is rarely tested with a complete three-way comparison (random initialization, SSL, and standard supervised transfer learning) across a full labeled-data efficiency curve on a single small, domain-specific dataset.

This is not merely a methodological gap. Whether SSL pretraining helps in the low-label regime depends on how much exploitable structure a dataset already contains relative to its size, and this can vary substantially across imaging modalities and even across individual datasets within the same modality. A pretraining method validated on one medical imaging benchmark may not transfer its reported advantage to another, and practitioners deciding how to allocate limited annotation budgets need evidence specific to their setting, not just an extrapolated intuition from natural-image SSL research.

In this work, we make the following contributions:

1. We present a complete three-way comparison , random initialization, SimCLR self-supervised pretraining, and ImageNet-supervised transfer learning , evaluated under both linear probing and full fine-tuning, across a full labeled-data efficiency curve (10%, 25%, 50%, 100%) on the BUSI breast ultrasound dataset.
2. We report a finding that runs counter to the common low-label SSL narrative: on BUSI, random-initialization and ImageNet-supervised fine-tuning both saturate almost immediately and remain essentially flat across the entire labeled-data range we test, whereas SimCLR is the only method that shows a meaningful, *rising* label-efficiency curve , making it the most, not least, label-hungry method in our comparison. We discuss why this is a plausible, dataset-specific outcome rather than an implementation failure.
3. We provide qualitative Grad-CAM [10] analysis comparing attention patterns across random-init, SimCLR, and ImageNet-supervised fine-tuned models, to assess whether quantitative performance differences correspond to differences in what each model attends to.
4. We release a fully reproducible, publicly available codebase and pretrained checkpoints, including all baseline and SSL training configurations, to support further investigation of when SSL pretraining does and does not provide a practical advantage in small medical imaging datasets.

2 Related Work

2.1 Deep Learning for Breast Ultrasound

Convolutional neural networks, including ResNet [7] and DenseNet [8] architectures, have been widely applied to breast ultrasound classification and lesion segmentation tasks [1]. More recently, Vision Transformers and Swin Transformer variants [9] have been explored for breast imaging, often in ensemble configurations with CNNs to combine local texture sensitivity with global context modeling. These approaches are predominantly supervised and therefore inherit the annotation bottleneck inherent to medical imaging.

2.2 Self-Supervised Learning

Contrastive self-supervised methods, including SimCLR [2], MoCo [5], and BYOL [6], learn representations by enforcing invariance to data augmentations across positive pairs while repelling negative pairs (or, in BYOL's case, without explicit negatives). These methods have demonstrated strong transfer performance on natural image benchmarks, typically evaluated via linear probing or fine-tuning on ImageNet or similarly large downstream datasets. Their behavior on small, domain-specific datasets with only a few hundred images , where the pretraining corpus and the downstream dataset are often the same small pool of images , is comparatively understudied, and the dynamics that govern SSL's effectiveness at this scale (e.g., overfitting of the pretraining objective itself, limited augmentation diversity, dataset-specific shortcut signal) differ substantially from the large-scale regime in which these methods were originally developed.

2.3 Self-Supervised Learning in Medical Imaging

Prior work has applied SSL to chest X-ray [11] and histopathology [12], generally reporting that SSL pretraining narrows the gap to supervised baselines, particularly in low-label regimes. These results, however, come from datasets that are substantially larger than BUSI (often tens of thousands of images), and it is not established whether the same low-label advantage holds at the much smaller scale typical of public breast ultrasound benchmarks. To our knowledge, no prior work has systematically evaluated a full random-init / SSL / supervised-transfer comparison across a labeled-data efficiency curve specifically on breast ultrasound , the gap our work addresses, with a result that complicates rather than confirms the standard narrative.

3 Method

3.1 Dataset

We use the publicly available Breast Ultrasound Images (BUSI) dataset [1], comprising 780 images across three classes: benign (437), malignant (210), and

normal (133). No additional private or institutional data is used, ensuring full reproducibility. Images are resized to 224×224 and converted to 3-channel format for compatibility with ImageNet-pretrained backbone initializations used in our supervised baseline. We use an 80/10/10 train/validation/test split, identical across all methods and label fractions, and 5-fold stratified cross-validation for all reported downstream results.

3.2 Self-Supervised Pretraining

We adopt the SimCLR v2 architecture [3]: a ResNet-50 encoder followed by a 2-layer MLP projection head ($2048 \to 2048 \to 128$) with batch normalization on the output layer to mitigate representational collapse. The model is trained with the NT-Xent contrastive loss:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where z_i, z_j are projected embeddings of two augmented views of the same image, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature hyperparameter, and N is the batch size. Pretraining uses only the unlabeled images from the training split (624 images); no held-out validation or test images are used during SSL pretraining, and labels are never used until the downstream evaluation stage.

Augmentations are adapted for ultrasound: we retain random resized cropping (scale 0.5–1.0, more conservative than the natural-image default of 0.08–1.0 to avoid excluding small lesion regions), horizontal flipping, color jitter, and Gaussian blur, but omit hue jitter, which is not meaningful for grayscale ultrasound and risks destroying speckle texture cues relevant to tissue characterization. SimCLR is trained for 200 epochs with a cosine learning rate schedule, 10-epoch linear warmup, an effective batch size of 256 (via gradient accumulation), the LARS optimizer, and temperature $\tau = 0.1$.

3.3 Downstream Evaluation Protocol

We evaluate the pretrained SimCLR encoder under two protocols: (1) *linear probing*, in which the encoder is frozen and only a linear classification head is trained, isolating representation quality; and (2) *full fine-tuning*, in which the entire network is updated, reflecting practical deployment performance. For both protocols, we vary the labeled training fraction ($10\%, 25\%, 50\%, 100\%$) to construct a labeled-data efficiency curve. All downstream results use stratified 5-fold cross-validation, and we report AUC, accuracy, sensitivity, and specificity as mean ± standard deviation across folds, computed once at the final training epoch of each fold (no checkpoint selection by validation performance, to avoid optimistic bias).

Table 1. Downstream classification performance on BUSI (mean ± std over 5 folds, final-epoch checkpoint, 100% labels). Best result per column in **bold**.

Method	AUC	Accuracy	Sensitivity	Specificity
Random init (supervised)	0.876 ± 0.022	0.744 ± 0.026	0.687 ± 0.033	0.847 ± 0.011
SimCLR (linear probe)	0.803 ± 0.036	0.665 ± 0.030	0.574 ± 0.038	0.788 ± 0.016
SimCLR (fine-tuned)	0.880 ± 0.033	0.744 ± 0.069	0.746 ± 0.047	0.864 ± 0.029
ImageNet supervised	**0.958 ± 0.010**	**0.860 ± 0.023**	**0.837 ± 0.037**	**0.914 ± 0.016**

3.4 Baselines

We compare against two baselines, evaluated under the identical 5-fold, label-fraction-matched protocol described above: (1) a *random-initialization* baseline, in which a ResNet-50 of the same architecture is trained from scratch with no pretraining of any kind; and (2) a *supervised ImageNet* baseline, in which the ResNet-50 is initialized from ImageNet-1k weights and fully fine-tuned. These two baselines bracket the comparison: random-init establishes what is achievable from the labeled data alone, while ImageNet-supervised establishes what a strong, widely used transfer-learning prior achieves under the same constraints.

4 Experiments and Results

4.1 Implementation Details

All models are trained and evaluated on a single consumer GPU (NVIDIA RTX 4050). SimCLR pretraining uses mixed-precision training and gradient accumulation to reach an effective batch size of 256. All experiments are logged via Weights & Biases for full training curve transparency. Code, configurations, and the exact data splits used are available at https://github.com/AZ1004/PULSE.

4.2 Main Results

Table 1 reports performance at 100% label availability across all four methods.

ImageNet-supervised fine-tuning achieves the highest AUC and accuracy by a substantial margin (0.958 vs. 0.880 for SimCLR fine-tuned, a gap of 7.8 AUC points that does not overlap within one standard deviation across folds). SimCLR fine-tuning consistently outperforms its own linear-probe variant (0.880 vs. 0.803 AUC), confirming that the pretrained representation, while not matching the ImageNet-supervised prior, contains non-trivial structure that benefits from end-to-end adaptation. Crucially, comparing these methods against the random-initialization floor highlights a stark boundary for local self-supervision on this benchmark. While SimCLR fine-tuning (0.880 ± 0.033 AUC) provides only a marginal benefit over training completely from scratch (0.876 ± 0.022 AUC), it fails to substantially alter downstream classification capacity. This indicates

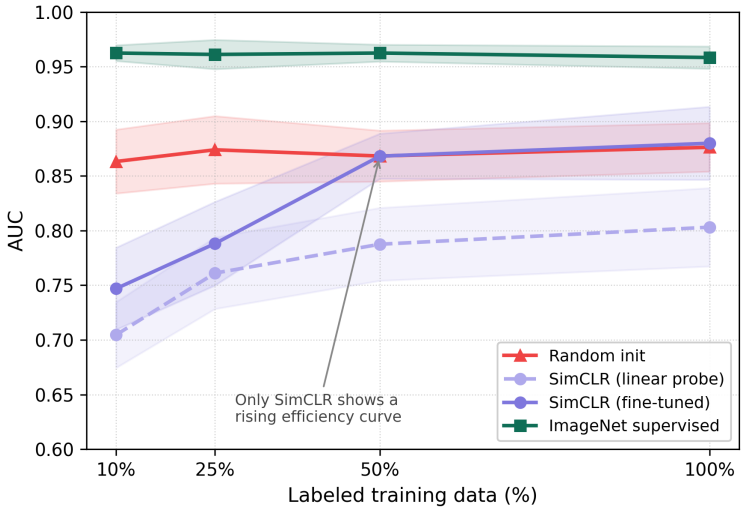

Fig. 1. Downstream AUC as a function of labeled training data fraction (10–100%). Shaded regions denote ±1 std across 5 folds. ImageNet-supervised and random-init are both approximately flat; SimCLR is the only method exhibiting a rising label-efficiency curve.

that contrastive instance discrimination, in this configuration, struggles to encode features that are materially superior to standard supervised optimization on the target domain, while the ImageNet-supervised baseline establishes a dominant performance ceiling (0.958 ± 0.010 AUC) that remains unchallenged by either alternative.

4.3 Label Efficiency

Figure 1 shows AUC as a function of labeled training fraction for all methods. At 10% label availability, SimCLR fine-tuned reaches an AUC of 0.747, compared to 0.880 at 100% labels , retaining 85% of its full-label performance with only a tenth of the labels, consistent with the premise that SSL pretraining provides most of its benefit even with very limited labeled data *relative to its own full-label ceiling*. However, this same pattern holds, and is in fact stronger, for the random-initialization baseline: at 10% labels (approximately 62 training images per fold), random initialization already reaches AUC 0.863 ± 0.029 , higher than SimCLR fine-tuned at the identical label fraction (0.747 ± 0.038). This pattern is even more pronounced for the ImageNet-supervised transfer learning baseline. At just 10% label availability (∼62 images), the ImageNet-initialized model achieves a mean macro AUC of 0.962 ± 0.007, matching its full-label performance within the margin of standard error. A rich natural-image prior thus appears to remove the data-scarcity constraint almost entirely on this benchmark.

4.4 Qualitative Analysis: Grad-CAM

Figure 2 visualizes Grad-CAM [10] attention maps for representative malignant cases across the random-init, SimCLR fine-tuned, and ImageNet-supervised fine-tuned models. Attention localization quality does not cleanly track the quantitative AUC ordering. Random-init and SimCLR fine-tuned produce visually similar attention patterns across samples , a broad region of high activation that substantially overlaps the lesion but frequently extends well beyond its boundary into surrounding tissue, with a notably consistent overall shape across images with quite different lesion morphologies. This consistency is itself informative: it suggests both models may be partly responding to a coarse positional prior (e.g., lesions tend to appear in a particular region of the image) rather than purely to lesion-specific texture.

The ImageNet-supervised model's attention is qualitatively distinct and, in our view, more concerning: in every sample we inspected, its strongest activation is concentrated near the right edge of the image, an area that frequently coincides with scanner-embedded text, depth markers, or other acquisition artifacts rather than the lesion itself, and this edge-locked pattern persists regardless of where the actual lesion is located within the frame. This raises the possibility that part of ImageNet-supervised's quantitative advantage reflects sensitivity to low-level acquisition artifacts in addition to, or instead of, lesion morphology , a hypothesis we investigate further below.

Annotation-reliance check. Motivated by the edge-concentrated attention observed for ImageNet-supervised, and by the fact that BUSI images occasionally contain scanner-embedded markers and, more notably, radiologist-added measurement annotations (calipers, bounding boxes, and text) that we observed to appear disproportionately on malignant images, we ran a small manual ablation. We inspected 20 randomly sampled malignant images and identified those with a visible annotation overlay; 2 of 20 had one. For each, we masked the annotated region (replacing it with the per-channel normalization mean, a neutral fill) and re-ran inference with the ImageNet-supervised model without retraining. The predicted probability of malignancy changed only modestly in both cases (mean $\Delta P(\text{malignant}) = -0.0195$, std 0.0011), and neither case crossed the classification threshold. While limited in scale, this check suggests that visible annotation artifacts, when present, are not the dominant driver of this model's malignant predictions, though they may contribute a small amount of additional, clinically irrelevant signal. We report this as a partial, not exhaustive, robustness check; a systematic, large-scale version of this test is an important direction for future work and is discussed further in Section 5.

5 Discussion and Limitations

Our results do not support the common expectation that self-supervised pre-training provides its largest relative advantage in the lowest-label regime, at

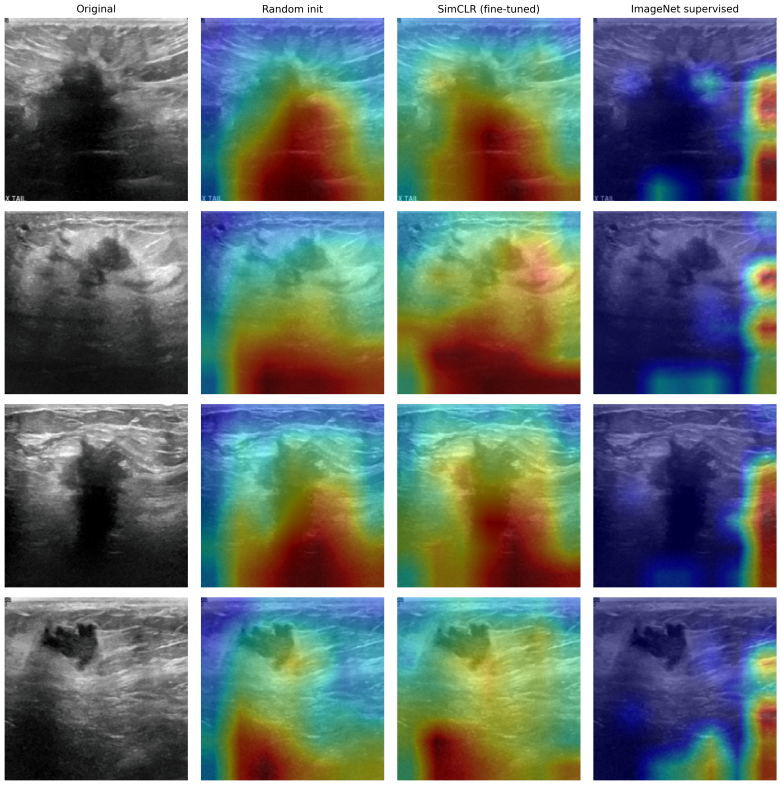

Fig. 2. Grad-CAM attention maps on representative malignant BUSI samples, comparing random initialization, SimCLR fine-tuned, and ImageNet-supervised fine-tuned encoders.

least on BUSI. Instead, we find that a randomly initialized encoder, trained from scratch on as few as 62 labeled images, already achieves a strong AUC (0.863), exceeding SimCLR fine-tuned at the same label fraction, and that SimCLR fine-tuning provides only a marginal improvement over random initialization even at full label availability (0.880 vs. 0.876 AUC). We consider this a genuine and informative negative result rather than an implementation shortcoming: BUSI images plausibly contain strong, low-level, class-correlated visual signal , texture, lesion shape regularity, acoustic shadowing patterns, or even acquisition-related artifacts , that a convolutional network can exploit with very little data and heavy augmentation, independent of any pretraining strategy. This is consistent with the observation of Zhang et al. [13], who find that contrastive objectives often optimize for coarse global structure rather than the fine-grained local features medical tasks require. Under this hypothesis, the binding constraint on this dataset is not representation learning (which random init already accomplishes adequately) but rather the strength of the prior encoded by ImageNet

pretraining, which SimCLR pretraining on only 624 ultrasound images cannot match.

Our Grad-CAM analysis and the accompanying annotation-reliance check (Section 4.4) add an important qualification to this picture: part of ImageNet-supervised's quantitative advantage may be attributable to sensitivity to image-border acquisition artifacts rather than lesion morphology alone, and BUSI's malignant images disproportionately, though not universally, carry visible radiologist annotations that could in principle serve as an unintended shortcut. Our small-scale check suggests these annotations are not the dominant driver of the model's predictions, but we did not test this exhaustively, and we flag annotation artifacts as a dataset-level confound that future work on BUSI , using any method, not only the ones evaluated here , should explicitly control for, for example by detecting and masking annotated regions during preprocessing. We note that the strong performance of random initialisation at very low label fractions (AUC 0.863 on 62 images) is itself consistent with shortcut exploitation rather than genuine lesion representation, and we cannot fully disentangle intrinsic low-level class-discriminative signal from dataset-level confounds on the basis of our small-scale annotation check alone. These two explanations, rich intrinsic signal and class-correlated artifacts, are not mutually exclusive, and distinguishing them would require either a systematic artifact-masking preprocessing step applied to all images, or replication on a dataset with controlled acquisition conditions.

This interpretation is consistent with, though distinct from, prior SSL findings on chest X-ray and histopathology [11,12], which typically pretrain on far larger unlabeled corpora (often combining multiple datasets, reaching tens of thousands of images) than the 624 images available for SimCLR pretraining here. Our results suggest that the reported low-label advantage of SSL in medical imaging may be conditional on pretraining corpus size in ways that are not yet well characterized, and that practitioners should not assume SSL pretraining will outperform standard supervised baselines by default on small, single-dataset medical imaging benchmarks.

Several limitations merit discussion. First, all results in this paper are specific to BUSI, a relatively small, single-center dataset collected under a single acquisition protocol. Our findings should not be assumed to generalise to other breast ultrasound datasets with different scanner types, patient demographics, or annotation practices. Second, we evaluate only SimCLR; other contrastive methods (MoCo, BYOL) and generative paradigms (MAE) may behave differently at this data scale, and pretraining on a larger pooled ultrasound corpus is a natural next step that may change the comparison.

6 Conclusion

We presented a complete random-init / self-supervised / supervised-transfer comparison for breast ultrasound classification across a labeled-data efficiency curve, testing the common assumption that SSL pretraining is most advan-

tageous when labels are scarce. On the BUSI dataset, we find the opposite of this assumption: random-initialization and ImageNet-supervised fine-tuning both saturate almost immediately and remain flat across the full label range we test, while SimCLR pretraining is the only method that requires substantially more labeled data to approach its own, lower ceiling. We interpret this as evidence that BUSI contains exploitable low-level signal accessible to standard supervised learning even with minimal data, while SimCLR's contrastive, augmentation-invariant features need labeled supervision to be steered toward the clinical decision boundary rather than simply read out. We release all code, configurations, and pretrained checkpoints to support further investigation of when , and when not , self-supervised pretraining is worth the additional training cost in label-scarce medical imaging settings. These conclusions are specific to the BUSI dataset and SimCLR; we caution against extrapolating them to other breast ultrasound benchmarks, acquisition settings, or self-supervised paradigms without further validation.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
3. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. In: NeurIPS (2020)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
5. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
6. Grill, J.B., Strub, F., Altché, F., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
9. Liu, Z., Lin, Y., Cao, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
10. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
11. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. In: MIDL (2021)
12. Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications **7**, 100198 (2022)
13. Zhang, C., Zheng, H., Gu, Y.: Dive into the details of self-supervised learning for medical image analysis. Medical Image Analysis **89**, 102879 (2023)