Benchmarking Self-Supervised Pretraining Strategies for Chest X-Ray Triage and Out-of-Distribution Transfer

Paola Martínez-Arias[1,2*], Begoña García-Zapirain Soto[2], Ibon Oleagordia Ruiz[2], and Karen López-Linares[1,3]

[1] Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastian (Spain)
[2] Faculty of Engineering, University of Deusto, Av. Universidades, 24, 48007 Bilbao (Spain)
[3] eHealth Group, Bioengineering Area, Biogipuzkoa Health Research Institute, Donostia-San Sebastian (Spain)
pnmartinez@vicomtech.org

Abstract. Self-supervised learning (SSL) has emerged as a powerful paradigm for medical image representation learning, reducing dependence on expert annotated datasets. However, a controlled comparison of the most prominent SSL families (reconstructive, contrastive, and hybrid) for the chest X-ray domain remains lacking. Thus, in this work, we benchmark Masked Autoencoders (MAE), Momentum Contrast v3 (MoCo v3), and Contrastive Masked Autoencoders (CMAE) by pretraining on MIMIC-CXR-JPG and evaluating on a chest X-ray triage task. To the best of our knowledge, this is the first adaptation of CMAE to this radiological domain. We evaluate representation quality through (i) linear probing with 100% labels, (ii) fine-tuning at 1%, 10%, and 100% label fractions on CheXpert, reporting AUROC and $F1_{macro}$ across multiple random seeds, and (iii) out-of-distribution (OOD) transfer in a few-shot regime. Our results show that MAE achieves the most consistent AUROC performance, with the clearest advantage at 10% labels and cross-anatomy OOD transfer. MoCo v3 shows a localized advantage in $F1_{macro}$ after post-hoc threshold selection, under the 1% label regime while all methods converge under full supervision. OOD transfer was weak overall, with MAE leading under cross-anatomy transfer, while the contrastive branch showed potential under cross-institution shift. The findings suggest that the choice of SSL strategy should be guided by the intended deployment scenario.

Keywords: Self-supervised learning · Chest X-ray · Contrastive learning · MAE · Representation learning · MIMIC-CXR-JPG · CheXpert

1 Introduction

Deep learning models for chest X-ray analysis have achieved remarkable performance in detecting a wide range of thoracic pathologies [14,23]. However, their

success has been conditioned on large collections of expert-labelled images, which are expensive and time-consuming to curate [1]. Self-supervised learning (SSL) alleviates this bottleneck by exploiting large amounts of unlabelled data to learn transferable representations prior to fine-tuning on downstream tasks [27].

Two dominant SSL families have emerged in the computer-vision community and subsequently adapted to the medical imaging field. **Contrastive learning methods** (MoCo [11], SimCLR [4]), learn instance-discriminative representations by pulling together augmented versions of the same image while pushing apart views from different images. Instead, **reconstructive methods** pioneered by the Masked Autoencoder (MAE) [10], learn to reconstruct masked image patches, encouraging the model to capture complete and locally consistent features. More recently, **hybrid approaches** such as Contrastive Masked Autoencoders (CMAE) [13], have emerged. They combine both objectives and have achieved state-of-the-art results on natural image benchmarks.

These paradigms have been applied to medical image analysis in prior work. On the contrastive side, MoCo-CXR [21] adapted the MoCo strategy to chest X-rays, while Azizi et al. [1] and Vu et al. [22] refined positive-pair construction using patient-level and structured metadata. On the reconstructive side, MAEs [10] have been adapted to medical images [29], including chest X-rays [25]. However, in the case of CMAE [13], even if MAE-based hybrid objectives have already been explored in medical vision-language settings [6], to our knowledge CMAE has not been applied to chest X-rays under controlled, single-modality conditions.

Despite progress on the area, controlled cross-family comparisons remain rare. A systematic review of several medical SSL studies [12] found that most of them compare a single strategy to a supervised baseline. Wolf et al. [24] compare contrastive and masked-autoencoder pretraining on small medical datasets, while Zheng et al. [28] benchmark contrastive, generative, and hybrid methods such as CMAE, but for CT segmentation rather than radiograph classification. This heterogeneity in SSL family, pretraining data, and training control makes it difficult to attribute performance differences to the SSL objective itself rather than to training scale or data distribution, and none of these studies jointly evaluates label efficiency and OOD transfer under matched pretraining conditions.

Since the benefits of self-supervised pretraining over training from scratch have already been established for chest X-ray classification [1,21,24], our objective is instead to characterize the relative behavior of reconstructive, contrastive, and hybrid SSL objectives under matched pretraining conditions, considering deployment-oriented scenarios in which models are adapted or calibrated using limited target-population data. Hereby, we address this gap by conducting a fair evaluation of MAE, MoCo v3, and CMAE pretrained on the same dataset (MIMIC-CXR-JPG [15]) and using the same backbone architecture (ViT-B/16 [8]). Our main contributions are:

1. We adapt and evaluate CMAE, previously unexplored in chest X-ray imaging.
2. We benchmark MAE, MoCo v3, and CMAE under matched pretraining conditions.

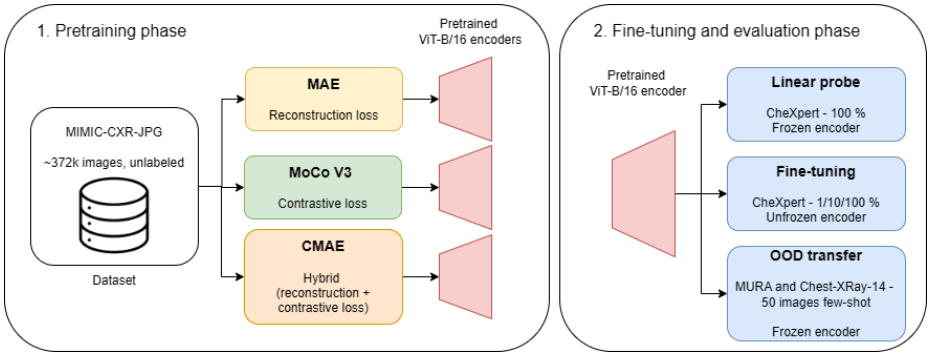

Fig. 1. Overview of the experimental pipeline. **Left**: three SSL methods (MAE, MoCo v3, CMAE) are pretrained from scratch on MIMIC-CXR-JPG, each producing a ViT-B/16 encoder. **Right**: each pretrained encoder is evaluated via linear probing, supervised fine-tuning at multiple label fractions, and OOD few-shot transfer.

3. We assess representational quality through linear probing, fine-tuning across label fractions (1%, 10%, 100%), and OOD few-shot transfer in cross-anatomy and cross-institution scenarios, for a comprehensive characterization of generalization, as performed in other benchmark comparisons [3].

2 Experimental setup and methodology

The methodology is divided into two stages, summarized in Figure 1: i) model pretraining on MIMIC-CXR-JPG using the 3 SSL strategies, and ii) evaluation and fine-tuning of the pretrained models with different amounts of labels using CheXpert, MURA/MSK and NIH-Chest-Xray14 (CXray14) for OOD analysis. Exact dataset sizes, label fractions, and number of independent repetitions are reported in Table 1. Protocols requiring subsampling were repeated with three random seeds, whereas full-data protocols were run once.

2.1 Datasets

MIMIC-CXR-JPG [15]. It is a publicly available collection of 377,110 chest X-rays from 227,835 studies from Beth Israel Deaconess Medical Center. We used ≈372k images for pretraining, and ≈5k images for validation, using the original splits to avoid data leakage between sets. Labels are not used at any stage of pretraining.

CheXpert [14]. This dataset contains 224,316 chest X-rays from 65,240 patients collected at Stanford University Medical Center, labelled across 14 thoracic conditions [14]. We binarize this into a *finding vs. no-finding* for our triage scenario. Using CheXbert-derived labels [20], a study is labelled *finding* if at least one of 13 pathologies is predicted present (excluding *No Finding*), with

Table 1. Summary of evaluation protocols. Seeds indicate number of independent repetitions

Protocol	Dataset	Label fraction	Metric	Seeds
Linear probe	CheXpert	100% (frozen enc.)	AUROC + $F1_{macro}$	1
Fine-tuning 1%	CheXpert	≈1,097 imgs	AUROC + $F1_{macro}$	3
Fine-tuning 10%	CheXpert	≈10,974 imgs	AUROC + $F1_{macro}$	3
Fine-tuning 100%	CheXpert	≈109,740 imgs	AUROC + $F1_{macro}$	1
OOD (cross-anatomy)	MSK/MURA	50 train, 3196 test	AUROC + $F1_{macro}$	3
OOD (cross-institution)	CXray14	50 train, 25596 test	AUROC + $F1_{macro}$	3

uncertain labels treated as positive to maximize sensitivity [26]. We preserve the official splits, undersampling *No Finding* cases in training to obtain a balanced set of 109,740 X-rays; validation (234) and test (668), each annotated by consensus of radiologists, follow the same binarization of labels and all samples are retained.

To evaluate cross-domain transfer we used the following datasets. In both cases, each original validation set was used as the test set:

MSK / MURA [19] tests cross-anatomy transfer. It is a large musculoskeletal X-ray dataset containing 40,561 images across seven body parts, evaluated on the normal/abnormal binary task.

CXray14 [23] tests cross-institution transfer within the same imaging modality. It comprises 112,120 frontal chest X-rays from the NIH Clinical Center, annotated with 14 disease labels. We binarize these into the same *finding vs. no-finding* scheme used for CheXpert.

2.2 SSL Pretraining

Setup: MAE [10], MoCo v3 [5] and CMAE [13] follow their official implementations. Augmentations that are not adequate for medical imaging such as color jitter and random grayscale are removed, and rotation and horizontal flipping were added, following recommendations in [21]. Images are resized to 224×224 pixels and normalized with dataset-specific mean and standard deviation computed from MIMIC-CXR-JPG ($\mu = 0.5020$ and $\sigma = 0.085585$).

Objectives and Augmentation Strategies: All models share the following configuration to ensure fair comparison. The backbone is a ViT-B/16 [8] with patch size 16×16 and input resolution 224×224. Pretraining is performed for 200 epochs on MIMIC-CXR-JPG [15]. We use the AdamW optimizer with cosine learning rate decay and a linear warm-up of 20 epochs. The effective batch size (EBS) is set to 1024 for the 3 models. Training is conducted using mixed-precision (bfloat16). Moreover, following the recommendations of the original implementations of MAE [10], MoCo v3 [5] and CMAE [13] we adopt a learning rate of 1.5×10^{-4}. Weight decay was set to 0.05 for all three models.

Table 2 summarizes the architecture, masking strategy, loss, and augmentation pipeline used for each SSL method. Unless otherwise noted, all hyperparameters follow the values recommended in each method's original publication rather than a domain-specific search.

Table 2. Configuration of the three SSL pretraining strategies following original implementations.

Method	Architecture	Masking/ Momentum	Loss	Augmentations
MAE [10]	Encoder–decoder (lightweight decoder)	Patch masking: 75%	Reconstruction loss $\mathcal{L}_{\text{rec}}$ using MSE on masked patches, normalized pixel space	– Random resized crop (0.7–1.0) – Horizontal flip
MoCo v3 [5]	Momentum encoder, 3-layer MLP projector, 2-layer predictor	Momentum $m = 0.99$	Contrastive loss $\mathcal{L}_{\text{cl}}$ using Symmetric InfoNCE between different views	– Random resized crop (0.7–1.0) – Horizontal flip – Gaussian blur – Rotation ($\pm 10°$)
CMAE [13]	Dual-branch: online encoder–decoder and momentum encoder. Additional feature decoder	*Online branch:* – Patch masking 65% *Momentum branch*: – momentum $m = 0.99$	Sum of reconstruction and contrastive loss $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{cl}}$, with $\lambda = 1.0$ following the original ablation study recommendation [13]	*Online branch:* – Random resized crop (0.7–1.0) – Horizontal flip *Momentum branch:* – Random resized crop (0.7–1.0) – Horizontal flip – Rotation ($\pm 10°$) – Gaussian blur – Pixel-shift

2.3 Finetuning and evaluation protocol

We evaluate the pretrained encoders under the three protocols shown in Figure 1 and detailed in Table 1. Models were trained for up to 50 epochs with AdamW and a weight decay of 0.05. Early stopping was applied using validation AUROC with a patience of 10 epochs, and the best checkpoint was used for test-set evaluation. The classification head learning rate was set to 1e-3 for all cases.

Linear probing: With the encoder frozen, we train a single linear head on 100% of the CheXpert training data to assess raw representation quality without adapting the backbone. We used a batch size of 256.

Supervised fine-tuning: We fine-tune the full encoder plus a linear head at 1%, 10%, and 100% label fractions. The 1% and 10% settings are repeated across 3 random seeds to estimate variance under limited supervision (each seed determines a different random subset of the training data), and the 100% setting, using the full training set, is run once. We use separate learning rates for the head and the backbone (1e-4 for the backbone, scaled to the EBS of 128).

OOD transfer: To assess cross-domain generalization without target domain backbone fine-tuning, we evaluate a frozen linear probe on a few-shot setup across 3 random seeds, under two domain-shift scenarios: cross-anatomy transfer to MSK/MURA and cross-institution transfer to CXray14. For each seed and each target dataset, 50 target-domain training images were sampled from the respective official training partition, stratified by class and anatomical region for MURA, and finding label for CXray14. When multiple images from the same patient or study were available, sampling was performed at the patient/study level to avoid leakage. We used a batch size of 16.

Performance was evaluated using AUROC and class-balanced F1_{macro}. AUROC was the primary metric because it is threshold-independent and measures

Table 3. Linear probe and fine-tuning (FT) performance on CheXpert (mean ± std across 3 seeds for the subsampled protocols; linear probe and FT 100% use the full training set and are therefore reported as single runs). $F1_{macro}$ calibration thresholds ranged 0.22–0.41 across methods and label fractions.

Method	Linear Probe		FT 1% labels		FT 10% labels		FT 100% labels	
	AUROC	$F1_{macro}$	AUROC	$F1_{macro}$	AUROC	$F1_{macro}$	AUROC	$F1_{macro}$
MAE	0.63	0.56	0.60 ± 0.09	0.54 ± 0.06	0.67 ± 0.004	0.58 ± 0.02	0.74	0.65
MoCo v3	0.63	0.58	0.59 ± 0.05	0.58 ± 0.04	0.60 ± 0.018	0.56 ± 0.02	0.74	0.65
CMAE	0.58	0.54	0.56 ± 0.03	0.52 ± 0.05	0.62 ± 0.003	0.56 ± 0.01	0.74	0.63

ranking ability for case prioritization. $F1_{macro}$ was used as a secondary operating-point metric, as binary triage requires threshold selection and a fixed threshold of 0.5 is inappropriate under the prevalence shift between the balanced training and test set (84% positive cases for CheXpert).

Thresholds were calibrated under a minimum recall constraint of $\geq 80\%$, chosen as a standardized sensitivity-oriented operating point for comparing methods in a triage-like scenario rather than a universal clinical deployment threshold. In practice, the appropriate threshold should depend on the workflow and be calibrated using expected target prevalence or a representative local cohort. To approximate this setting, we estimated $F1_{macro}$ using repeated stratified 5-fold cross-validation within the held-out target cohort, so that no sample was used to determine its own threshold. In each fold, the threshold was selected on the remaining folds as the most selective value achieving recall $\geq 80\%$, and then applied to the held-out fold. Thus, $F1_{macro}$ represents an out-of-fold estimate after post-hoc calibration, while AUROC remains the primary metric for model comparison. Statistical significance is assessed using pairwise DeLong's tests for AUROC [7]. $F1_{macro}$ differences, as well as OOD AUROC comparisons against chance level, are assessed using patient-clustered bootstrap resampling with 2000 iterations [9]. For the three pairwise model comparisons within each protocol, we apply Benjamini–Hochberg (BH) correction for multiple settings [2].

3 Results and Discussion

3.1 Linear probe and fine-tuning

Table 3 reports linear probing and fine-tuning results across label fractions. At linear probing, MAE and MoCo v3 achieve the highest AUROC (0.63) and are statistically indistinguishable (DeLong, $p_{BH} = 0.57$), while both significantly outperform CMAE (0.58; $p_{BH} = 0.003$ and 0.020 respectively), suggesting that CMAE's frozen representations are less globally discriminative in this regime. Prior work [10,17] suggests that reconstruction-based objectives tend to produce rich but less linearly separable features than contrastive methods; MAE deviates from this pattern here, possibly due to the narrower domain of chest X-rays and the binary nature of the downstream task, making features easier to separate.

Table 4. OOD transfer under two domain-shift scenarios using few-shot: cross-anatomy transfer to MSK/MURA and cross-institution transfer to CXray14 (mean ± std across 3 seeds). $F1_{macro}$ calibrated to 80% recall as in Table 3.

Method	Cross-anatomy (MURA)		Cross-institution (CXray14)	
	AUROC	$F1_{macro}$	AUROC	$F1_{macro}$
MAE	0.587 ± 0.018	0.498 ± 0.005	0.587 ± 0.042	0.551 ± 0.049
MoCo v3	0.535 ± 0.044	0.481 ± 0.041	0.594 ± 0.009	0.556 ± 0.010
CMAE	0.567 ± 0.052	0.495 ± 0.062	0.598 ± 0.017	0.562 ± 0.025

CMAE, in contrast, may be subject to a trade-off as the encoder may not fully optimize global discriminability while simultaneously satisfying the reconstruction objective. MoCo v3's competitive performance supports the strong linear separability typically associated with contrastive representations [4,11].

At the **1% label** fraction, all three methods perform comparably (AUROC between 0.56 and 0.60) with overlapping standard deviations, and are statistically indistinguishable. This indicates that the available supervision is insufficient to reliably differentiate pretraining strategies at this extreme low-data regime. However, considering $F1_{macro}$, MoCo v3 (0.58) significantly outperforms both CMAE (0.52, $p_{BH} = 0.0056$) and MAE (0.54, $p_{BH} = 0.035$) at the calibrated operating point, suggesting a better precision–recall trade-off under the selected 80% recall constraint. At **10% labels**, differences begin to emerge: MAE achieves significantly higher AUROC (0.67), followed by CMAE (0.62) and MoCo v3 (0.60). In terms of $F1_{macro}$, differences do not reach significance after correction ($p_{BH} \geq 0.69$). At **100% labels**, AUROC converges to the same value (0.74) for all three methods, with no significant pairwise differences ($p_{BH} \geq 0.91$), indicating that sufficient supervision closes any representational gap between pretraining strategies. $F1_{macro}$ likewise shows no significant differences at this fraction (MAE 0.65, MoCo v3 0.65, CMAE 0.63).

3.2 OOD transfer:

Table 4 presents few-shot OOD transfer results under two domain shift scenarios: cross-anatomy results from chest X-rays (pretrained on MIMIC-CXR-JPG) to musculoskeletal X-rays (MSK/MURA), and cross-institution transfer to chest X-rays from a different source (CXray14).

Cross-anatomy transfer (MURA). Performance is weak overall but not uniformly at chance level: bootstrap testing against chance (AUROC = 0.5) shows MAE is significantly above chance in all three seeds ($p \leq 0.002$), CMAE in two of three seeds, and MoCo v3 in only one of three seeds. Pairwise comparisons confirmed the AUROC ranking MAE > CMAE > MoCo v3 after BH correction ($p_{BH} \leq 0.029$). $F1_{macro}$ showed the same ordering and was also statistically significant after correction. However, the absolute differences were small. We therefore interpret the OOD results as statistically consistent but practically modest.

Cross-institution transfer (CXray14). Unlike cross-anatomy transfer, all three methods perform significantly above chance in all three seeds ($p = 0.001$ in every case), indicating that some discriminative signal is retained by every SSL objective under institutional shift alone. However, the ranking among methods differs from the cross-anatomy setting: CMAE achieves the highest mean AUROC (0.598), followed by MoCo v3 (0.594) and MAE (0.587). Pairwise AUROC differences were not statistically significant after BH correction ($p_{\mathrm{BH}} \geq 0.20$), so at the AUROC level the three objectives are statistically indistinguishable under this shift. $\mathrm{F1}_{\mathrm{macro}}$, in contrast, showed the same CMAE > MoCo v3 > MAE ordering and was statistically significant after correction ($p_{\mathrm{BH}} < 0.001$ for all three pairwise comparisons), indicating a modest advantage for CMAE at the calibrated operating point.

The two protocols do not show a consistent ranking across domain-shifts. MAE shows the clearest advantage under cross-anatomy transfer, whereas CMAE shows a modest advantage in $\mathrm{F1}_{\mathrm{macro}}$ under cross-institution transfer. In both cases, the leading method includes a reconstruction-based objective suggesting that the low-level structural features preserved through reconstruction, such as texture and contrast, may support cross-domain transfer. The contrastive branch in CMAE may provide an additional benefit under acquisition-level shifts, where anatomy is similar but scanner, protocol, or patient characteristics differ. However, given the small numerical difference and the limited few-shot target sets in both protocols, these results should be considered exploratory and validated on additional domains and larger few-shot settings.

3.3 Clinical use case: implications for triage

The observed pattern is relevant for deployment-oriented scenarios in which models are adapted or calibrated using limited target-population data. In particular, for triage-like settings where the preferred SSL strategy may depend on whether the goal is case prioritization or binary decision-making at a calibrated operating point. Under this framing, MAE appears more suitable when ranking performance is prioritized under label-scarce fine-tuning, whereas MoCo v3 shows an advantage for $\mathrm{F1}_{\mathrm{macro}}$ in the most label-scarce setting. We therefore interpret our results not as unconditional superiority of one strategy, but as a label-fraction-specific choice. Consequently, post-hoc threshold calibration [18,16] on the target deployment population should be considered necessary before clinical deployment regardless of pretraining strategy. The **OOD results** suggest that this choice may also depend on the intended distribution shift. In the cross-anatomy scenario, MAE shows a clear advantage indicating that reconstructive objectives may help retain transferable low-level features. However, in the cross-institution scenario which is more clinically relevant, reflecting deployment across institutions, CMAE shows a modest advantage in $\mathrm{F1}_{\mathrm{macro}}$, indicating that the addition of the contrastive branch may suit acquisition-level variability across sites. Nevertheless, the overall OOD performance remained weak for both scenarios under this minimal adaptation setting, showing that single-domain chest X-ray pretraining may be insufficient for robust generalization beyond the source distribution.

4 Conclusions

We presented a controlled benchmark of three SSL strategies pretrained from scratch on MIMIC-CXR-JPG for chest X-ray representation learning: MAE, MoCo v3, and CMAE, including, to the best of our knowledge, the first adaptation of CMAE to this domain. Across the evaluated regimes, MAE emerged as the most consistent strategy for ranking-based performance, with its clearest advantage at 10% labels and under cross-anatomy few-shot transfer to MSK/MURA. MoCo v3 showed a localized advantage in $F1_{macro}$ under the most label-scarce setting, but this difference did not persist with more labelled data. Under full supervision, all methods converged to similar performance. This pattern suggests that SSL method selection in practice should be guided by the intended deployment scenario and labelling budget rather than by a single aggregate metric. CMAE's hybrid objective does not consistently outperform the individual reconstruction or contrastive baselines, and does not surpass MAE within CheXpert fine-tuning regimes. However, it showed a modest advantage over MAE and MoCo v3 under cross-institution OOD transfer, suggesting that the benefits of combining reconstruction and contrastive objectives may depend not only on optimized settings, but also on the type of distribution shift.

Even if our preliminary results suggest that objectives involving reconstruction provide the most consistent representation across regimes, specific adjustment of hyperparameters, backbone size, and pretraining schedule, together with larger OOD and multi-label cohorts, were not explored here. In addition, contrastive methods are highly sensitive to augmentation design and batch size. We therefore leave a systematic study of the influence of these factors to future work.

Acknowledgments. This work is funded by the Basque Government under the HAZITEK 2024 programme (grant number ZE- 2024/00030).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azizi, S., Mustafa, B., Ryan, F., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3478–3488 (2021)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B **57**(1), 289–300 (1995)
3. Bommasani, R., Hudson, D.A., Adeli, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
4. Chen, T., Kornblith, S., Norouzi, M., et al.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)

5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9640–9649 (2021)
6. Chen, Z., Du, Y., Hu, J., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 679–689 (2022)
7. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics **44**(3), 837–845 (1988)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC (1993)
10. He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
11. He, K., Fan, H., Wu, Y., et al.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020)
12. Huang, S.C., Pareek, A., Jensen, M., et al.: Self-supervised learning for medical image classification: A systematic review and implementation guidelines. npj Digital Medicine **6**(1), 74 (2023). https://doi.org/10.1038/s41746-023-00811-0
13. Huang, Z., Jin, X., Lu, C., et al.: Contrastive masked autoencoders are stronger vision learners. arXiv preprint arXiv:2207.13532 (2022)
14. Irvin, J., Rajpurkar, P., Ko, M., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)
15. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019). https://doi.org/10.1038/s41597-019-0322-0
16. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML). pp. 625–632. ACM (2005)
17. Park, J.Y., Ayromlou, S., Geifman, A., et al.: A closer look at benchmarking self-supervised pre-training with image classification. International Journal of Computer Vision (2024)
18. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers. pp. 61–74. MIT Press (1999)
19. Rajpurkar, P., Irvin, J., Bagul, A., et al.: MURA: Large dataset for abnormality detection in musculoskeletal radiographs. In: Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL) (2018)
20. Smit, A., Mukherjee, S., Rajpurkar, P., et al.: CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020)
21. Sowrirajan, H., Yang, J., Ng, A.Y., et al.: MoCo pretraining improves representation and transferability of chest x-ray models. In: Proceedings of the Medical Imaging with Deep Learning (MIDL). pp. 728–744 (2021)

22. Vu, Y.N.T., Wang, R., Balachandar, N., et al.: MedAug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. In: Machine Learning for Healthcare Conference (MLHC). pp. 755–769 (2021)
23. Wang, X., Peng, Y., Lu, L., et al.: ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2097–2106 (2017)
24. Wolf, D., Payer, T., Lisson, C.S., et al.: Self-supervised pre-training with contrastive and masked autoencoder methods for dealing with small datasets in deep learning for medical imaging. Scientific Reports **13**(1), 20260 (2023)
25. Xiao, J., Bai, Y., Yuille, A., et al.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3588–3600 (2023)
26. Yuan, X., Liao, W., Peng, J., et al.: Large-scale chest x-ray text analysis and automatic chest x-ray report generation. arXiv preprint arXiv:2111.04917 (2021)
27. Zhang, C., Zhang, C., Song, J., et al.: A survey on masked autoencoder for self-supervised learning in vision and beyond. arXiv preprint arXiv:2208.00173 (2022)
28. Zheng, J., Wen, R., Hu, H., et al.: Tissue-contrastive semi-masked autoencoders for segmentation pretraining on chest CT. arXiv preprint arXiv:2407.08961 (2024)
29. Zhou, L., Liu, H., Bae, J., et al.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–6 (2023)