# Contrastive Masked Image-Text Modeling for Medical Visual Representation Learning

Cheng Chen[1], Aoxiao Zhong[2], Dufan Wu[1], Jie Luo[1], and Quanzheng Li[1,3(✉)]

[1] Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[2] School of Engineering and Applied Sciences, Harvard University, Boston, MA, USA
[3] Data Science Office, Massachusetts General Brigham, Boston, MA, USA
li.quanzheng@mgh.harvard.edu

**Abstract.** Self-supervised learning (SSL) of visual representations from paired medical images and text reports has recently shown great promise for various downstream tasks. However, previous work has focused on investigating the effectiveness of two major SSL techniques separately, i.e., contrastive learning and masked autoencoding, without exploring their potential synergies. In this paper, we aim to integrate the strengths of these two techniques by proposing a contrastive masked image-text modeling framework for medical visual representation learning. On one hand, our framework conducts cross-modal contrastive learning between masked medical images and text reports, with a representation decoder being incorporated to recover the misaligned information in the masked images. On the other hand, to further leverage masked autoencoding, a masked image is also required to be able to reconstruct the original image itself and the masked information in the text reports. With pre-training on a large-scale medical image and report dataset, our framework shows complementary benefits of integrating the two SSL techniques on four downstream classification datasets. Extensive evaluations demonstrate consistent improvements of our method over state-of-the-art approaches, especially when very scarce labeled data are available. code is available at https://github.com/cchen-cc/CMITM.

**Keywords:** Image-text representation learning · masked autoencoding · contrastive learning

## 1 Introduction

Deep learning models have demonstrated undoubted potential in achieving expert-level medical image interpretation when powered by large-scale labeled datasets [4,7,20]. However, medical image annotations require expert knowledge thus are extremely costly and difficult to obtain at scale. Such an issue has even become a bottleneck for advancing deep learning models in medical applications. To tackle this issue, recent efforts have resorted to the text reports that


H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 493–503, 2023.
https://doi.org/10.1007/978-3-031-43904-9_48

are paired with the medical images, aiming to leverage the detailed text interpretation provided by radiologists to assist the representation learning of medical images without relying on any manual labels [3,10,18,25–27]. The learned image representations have proven to be generalizable to other downstream tasks of medical image analysis, which can significantly reduce the amount of labeled data required for fine-tuning. This topic has been actively studied and seen rapid progress with evaluation on chest X-ray datasets, because of both the clinical importance of radiograph screening and the availability of large-scale datasets of public chest X-ray images with paired radiology reports [14].

The current mainstream approaches for medical image-text pre-training are based on the popular self-supervised learning (SSL) technique known as contrastive learning [2,9], which maximizes agreement between global and local representations of paired images and reports with a contrastive loss [1,10,15, 22,26,27]. These methods have demonstrated the effectiveness of using medical reports as a form of free supervision to enhance the learning of general image representations. Another well-demonstrated SSL method is masked autoencoding, which achieves representation learning via solving the pretext task of recovering masked image patches with unmasked ones [8]. Until very recently, the potential of masked autoencoding had only begun to be explored for medical image-text pre-training. In a latest work, Zhou et al. [28] propose to learn radiograph representations with a unified framework that requires the unmasked image patches to recover masked images and complete text reports. While both contrastive learning and masked autoencoding have demonstrated their ability to learn effective image representations, the two major SSL techniques have only been separately explored. The latest research has started to combine the two SSL techniques for joint benefits, but their focus is on the image domain rather than cross-modal learning in medical images and paired text reports [11,13,17]. It remains an interesting and unexplored question whether contrastive learning and masked autoencoding can benefit each other and how to jointly exploit their strengths for medical image-text pre-training.

In fact, the learning principles of contrastive learning and masked autoencoding suggest that they could be complementary to each other. Contrastive image-text learning explicitly discriminates the positive and negative pairs of images and text reports, making it good at promoting strong discriminative capabilities of image representations. Instead, masked autoencoding aims to reconstruct masked image/text tokens, which emphasizes learning local image structures, but may be less effective in capturing discriminative representations. This motivates us to propose a novel **c**ontrastive **m**asked **i**mage-**t**ext **m**odeling (CMITM) method for medical visual representation learning. Our framework is designed to accomplish three self-supervised learning tasks: First, aligning the representations of masked images with text reports. Second, reconstructing the masked images themselves. Third, reconstructing the masked text reports using the learned image representations. To reduce the information misalignment between the masked images and text reports, we incorporate a representation decoder to recover the missed information in images, which benefits the cross-modal learning. Moreover, the synergy of contrastive learning and masked autoencoding is

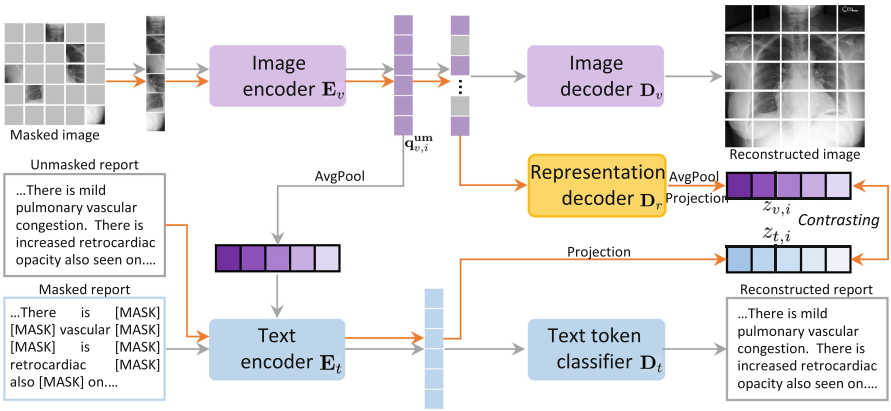


**Fig. 1.** Overview of proposed *contrastive masked image-text modeling* (CMITM) framework for medical visual representation learning. The masked image patches are required to align with the text reports, reconstructing original images, and reconstructing original text reports. The data flow in gray and orange line corresponds to first- and second-stage of pre-training respectively.

unleashed via a cascaded training strategy. Our framework is pre-trained on a large-scale medical dataset MIMIC-CXR with paired chest X-ray images and reports, and is extensively validated on four downstream classification datasets with improved fine-tuning performance. Combining the two techniques yields consistent performance increase and the improvement of our method even surpasses the benefits of adding data from 1% to 100% labels on CheXpert dataset.

## 2 Method

Figure 1 illustrates our contrastive masked image-text modeling (CMITM) framework. In this section, we first present how the cross-modal contrastive learning and masked autoencoding are realized in the framework respectively. Then we introduce the training procedures and implementation details of the framework.

### 2.1 Cross-Modal Contrastive Learning with Masked Images

Cross-modal contrastive learning has demonstrated to be an effective tool to align the representations of a medical image with that of its paired text report. In this way, the network is guided to interpret the image contents with the knowledge provided by medical reports. Different from previous methods, the cross-modal contrastive learning in our framework is between the representations of masked images and unmasked reports, aiming to integrate the benefits of both contrastive learning and masked image modeling.

Specifically, denote $\mathcal{D} = \{x_{v,i}, x_{t,i}\}_{i=1}^{N}$ as the multi-modal dataset consisting of $N$ pairs of medical images $x_{v,i}$ and medical reports $x_{t,i}$. Each input image

is split into $16 \times 16$ non-overlap patches and tokenized as image tokens $a_{v,i}$, and each text report is also tokenized as text tokens $a_{t,i}$. A random subset of image patches is masked out following masked autoencoder (MAE) [8]. As shown in Fig. 1, the unmasked patches are forwarded to the image encoder $\mathrm{E}_v$, which embeds the inputs by a linear projection layer with added positional embeddings and then applies a series of transformer blocks to obtain token representations of unmasked patches $\mathrm{q}^{\mathrm{um}}_{v,i}$. Directly utilizing the representations of only unmasked patches to perform contrastive learning with the text could be less effective, since a large portion of image contents has been masked out and the information from the images and texts are misaligned. To recover the missing information in the images, we feed both the encoded visible patches $\mathrm{q}^{\mathrm{um}}_{v,i}$ and trainable mask tokens $\mathrm{q}^{\mathrm{m}}_{v,i}$ with added positional embeddings $\mathrm{e}^{p}_{v,i}$ to a representation decoder $\mathrm{D}_r$ with two layers of transformer blocks. The representation decoder aims to output the representations of all image patches, i.e., $\hat{\mathrm{q}}_{v,i} = \mathrm{D}_r([\mathrm{q}^{\mathrm{um}}_{v,i}, \mathrm{q}^{\mathrm{m}}_{v,i}] + \mathrm{e}^{p}_{v,i})$. Such a design helps to avoid that the contrastive learning is confused by the misaligned information between masked images and text reports. Finally we apply a global average pooling operation and a project layer $\mathrm{h}_v$ to obtain the image embeddings $z_{v,i}$, i.e., $z_{v,i} = \mathrm{h}_v(\mathrm{AvgPool}(\hat{\mathrm{q}}_{v,i})$. For text branch, we consider the full reports could give more meaningful guidance to image understanding than masked ones. So we forward the full text tokens without masking to the text encoder $\mathrm{E}_t$ and the text project layer $\mathrm{h}_t$ to obtain the global text embeddings $z_{t,i}$, i.e., $z_{t,i} = \mathrm{h}_t(\mathrm{E}_t(a_{t,i}))$. To ensure that the embeddings of images are aligned with those of paired texts while remaining distant from unpaired texts, we employ cross-modal contrastive learning with the following symmetric InfoNCE loss [19].

$$\mathcal{L}_c = -\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(s^{vt}_{i,i}/\tau)}{\sum\exp(s^{vt}_{i,k}/\tau)} + \log\frac{\exp(s^{tv}_{i,i}/\tau)}{\sum\exp(s^{tv}_{i,k}/\tau)}\right], \tag{1}$$

where $s^{vt}_{i,j} = z^T_{v,i} z_{t,j}$, $s^{tv}_{i,j} = z^T_{t,i} z_{v,j}$, $\tau$ denotes the temperature which is set to be 0.07 following common practice, and $B$ is the number of image-report pairs in a batch. The cross-modal contrastive loss is used to supervise the network training associated with the data flow in orange line in Fig. 1.

### 2.2 Masked Image-Text Modeling

The masked image-text modeling component in our framework consists of two parallel tasks, i.e., masked image reconstruction with image only information and masked text reconstruction with cross-modal information. We follow the design in [8,28] for the masked image-text modeling since our main focus is whether masked autoencoding and contrastive learning can have joint benefits.

**Masked Image Reconstruction.** As aforementioned, the input images are masked and processed by the image encoder $\mathrm{E}_v$ to obtain $\mathrm{q}^{\mathrm{um}}_{v,i}$. As shown in Fig. 1, besides the representation decoder to reconstruct image representations, we also connect both the encoded visible patches and learnable unmasked tokens with added positional embeddings to an image decoder $\mathrm{D}_v$ to reconstruct the pixel values of masked patches, i.e., $\hat{x}_{v,i} = \mathrm{D}_v([\mathrm{q}^{\mathrm{um}}_{v,i}, \mathrm{q}^{\mathrm{m}}_{v,i}] + \mathrm{e}^{p}_{v,i}))$. The image decoder consists of a series of transformer blocks and a linear projection layer

to predict the values for each pixel in a patch. To enhance the learning of local details, the image decoder is required to reconstruct a high-resolution patch, which is twice the input resolution [28]. The training of image reconstruction is supervised by a mean squared error loss $\mathcal{L}_{vr}$ which computes the difference between the reconstructed and original pixel values for the masked patches:

$$\mathcal{L}_{vr} = \frac{1}{B}\sum_{i=1}^{B}\frac{\sum(\hat{x}^{\mathrm{m}}_{v,i} - \mathrm{norm}(\tilde{x}^{\mathrm{m}}_{v,i}))^2}{|\tilde{x}^{\mathrm{m}}_{v,i}|}, \tag{2}$$

where $\hat{x}^{\mathrm{m}}_{v,i}$, $\tilde{x}^{\mathrm{m}}_{v,i}$ denote the predicted and the original high-resolution masked patches, $|\cdot|$ calculates the number of masked patches, and norm denotes the pixel normalization with the mean and standard deviation of all pixels in a patch suggested in MAE [8]. The loss is only computed on the masked patches.

**Cross-Modal Masked Text Reconstruction.** To make the most of the text reports paired with imaging data for learning visual representations, the task of cross-modal masked text modeling aims to encourage the encoded visible image tokens $\mathrm{q}^{\mathrm{um}}_{v,i}$ to participate in completing the masked text reports. Specifically, besides the full texts, we also forward a masked text report with a masking ratio of 50% to the text encoder $\mathrm{E}_t$. Following [28], this value is deliberately set to be higher than the masking ratio of 15% in BERT [5] in order to enforce the image encoder to better understand the image contents by trying to reconstruct a large portion of masked texts. Then the global embedding of corresponding unmasked image patches $\mathrm{AvgPool}(\mathrm{q}^{\mathrm{um}}_{v,i})$ is added to the text token embeddings $\mathrm{q}^{\mathrm{um}}_{t,i}$ to form a multi-modal embeddings. To reconstruct the masked text tokens, the multi-modal embeddings are processed by the text encoder $\mathrm{E}_t$ and a text token classifier $\mathrm{D}_t$, i.e., $\hat{a}_{t,i} = \mathrm{D}_t(\mathrm{E}_t(\mathrm{q}^{\mathrm{um}}_{t,i} + \mathrm{AvgPool}(\mathrm{q}^{\mathrm{um}}_{v,i})))$. The training of text reconstruction is supervised by the cross entropy loss between the predictions and original text tokens as follows:

$$\mathcal{L}_{tr} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{H}(a^{\mathrm{m}}_{t,i},\ \hat{a}^{\mathrm{m}}_{t,i}), \tag{3}$$

where $a^{\mathrm{m}}_{t,i}$, $\hat{a}^{\mathrm{m}}_{t,i}$ denote the original and recovered masked text tokens respectively, $\mathcal{H}$ denotes the cross entropy loss. Similar to masked image reconstruction, the loss is also only computed on the masked text tokens. The image and text reconstruction losses are used to supervise the network training associated with the data flow in gray line in Fig. 1.

### 2.3 Training Procedures and Implementation Details

Training a framework that combines cross-modal contrastive learning and masked autoencoding is non-trivial. As observed in prior work [13], forming the training as a parallel multi-task learning task can lead to decreased performance, which might be caused by the conflicting gradients of the contrastive and reconstruction losses. Similar phenomenon has also been observed in our experiments. We therefore adopt a cascaded training strategy, that is the framework is first trained with the reconstruction loss $\mathcal{L}_r = \mathcal{L}_{vr} + \mathcal{L}_{tr}$ and is further trained with

the contrastive loss $\mathcal{L}_c$. Such a training order is considered based on the insights that masked autoencoding focuses more on the lower layers with local details while contrastive learning is effective in learning semantic information for higher layers. Specifically, the first stage of pre-training follows [28] to employ the loss $\mathcal{L}_r$ for training 200 epochs. The model is trained using AdamW [16] optimizer with a learning rate of 1.5e-4 and a weight decay of 0.05. In the second stage of pre-training, the image encoder, text encoder, and representation decoder are further trained with the loss $\mathcal{L}_c$ for 50 epochs. Similarly a AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.05 is adopted. The framework is implemented on 4 pieces of Tesla V100 GPU with a batch size of 256. For network configurations, we use ViT-B/16 [6] as the image encoder and BERT [5] as the text encoder. The image decoder and representation decoder consists of four transformer blocks and two transformer blocks respectively.

## 3 Experiments

To validate our framework, the image encoder of the pre-trained model is used to initialize a classification network with a ViT-B/16 backbone and a linear classification head. We adopt the fine-tuning strategy as used in [26,28], where both the encoder and classification head are fine-tuned. This fine-tuning setting reflects how the pre-trained weights can be applied in practical applications. For each dataset, the model is fine-tuned with 1%, 10%, and 100% labeled training data to extensively evaluate the data efficiency of different pre-trained models. The dataset split remains consistent across all approaches.

**Pre-training Dataset.** To pre-train our framework, we utilize **MIMIC-CXR** dataset [14], which is a large public chest X-ray collection. The dataset contains 377,110 images extracted from 227,835 radiographic studies and each radiograph is associated with one radiology report. For pre-training, images are resized and randomly cropped into the size of $448 \times 448$ and $224 \times 224$ as the high-resolution image reconstruction ground truth and low-resolution inputs respectively.

**Fine-Tuning Datasets.** We transfer the learned image representations to four datasets for chest X-ray classification. **NIH ChestX-ray** [24] includes 112,120 chest X-ray images with 14 disease labels. Each chest radiograph can associate with multiple diseases. We follow [28] to split the dataset into 70%/10%/20% for training, validation, and testing. **CheXpert** [12] comprises 191,229 chest X-ray images for multi-label classification, i.e., atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. We follow previous work to use the official validation set as the test images and randomly split 5,000 images from training data as the validation set. **RSNA** Pneumonia [21] dataset contains 29,684 chest X-rays for a binary classification task of distinguishing between normal and pneumonia. Following [28], the dataset is split as training/validation/testing with 25,184/1,500/3,000 images respectively. **COVIDx** [23] is a three-class classification dataset with 29,986 chest radiographs from 16,648 patients. The task

**Table 1.** Quantitative comparison with different pre-training methods on four chest X-ray datasets when fine-tuning with 1%, 10%, and 100% training data.

| Methods | NIH X-ray (AUC) | | | CheXpert (AUC) | | | RSNA (AUC) | | | COVIDx (ACC) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% |
| Random init | 60.0 | 65.2 | 72.8 | 70.4 | 81.1 | 85.8 | 71.9 | 82.2 | 88.5 | 64.2 | 75.4 | 87.7 |
| ImageNet init | 69.8 | 74.4 | 80.0 | 80.1 | 84.8 | 87.6 | 83.1 | 87.3 | 90.8 | 72.0 | 84.4 | 90.3 |
| MAE [8] | 74.7 | 81.3 | 85.1 | 80.7 | 86.0 | 86.7 | 84.2 | 89.6 | 91.3 | 69.8 | 82.3 | 90.7 |
| MRM [28] | 79.4 | 84.0 | 85.9 | 88.5 | 88.5 | 88.7 | 91.3 | **92.7** | 93.3 | 78.0 | 90.2 | 92.5 |
| GLoRIA [10] | 77.7 | 82.8 | 85.0 | 86.5 | 87.5 | 87.8 | 89.7 | 91.2 | 92.1 | 76.7 | **91.7** | 94.8 |
| MGCA [22] | 78.2 | 82.7 | 85.0 | 87.0 | 88.4 | 88.5 | 90.7 | 92.6 | **93.4** | 75.2 | 91.5 | 94.3 |
| CMITM (ours) | **80.4** | **84.1** | **86.0** | **89.0** | **89.0** | **89.2** | **91.6** | 92.6 | **93.4** | **79.5** | 90.2 | **95.3** |

is to classify each image into normal, COVID-19 and non-COVID pneumonia. Same as [22], we use the official validation data as the test data and split 10% images from the training data for validation. For fine-tuning, images are also resized and randomly cropped into the size of $224 \times 224$.

**Comparison with State-of-the-Art Methods.** We compare our method with four state-of-the-art SSL methods including two masked autoencoding-based methods and two contrastive learning-based methods. **MAE (CVPR 2022)** [8] is the representative work on masked image autoencoding. **MRM (ICLR 2023)** [28] is the latest work on medical image-text pre-training by using both the self- and report-completion objectives based on the masked record modeling. **GLoRIA (ICCV 2021)** [10] and **MGCA (NeurIPS 2022)** [22] are two cross-modal medical visual representation learning methods based on multi-scale image-text contrastive learning. For a fair comparison, the results of MRM model on the datasets NIH ChestX-ray, CheXpert, and RSNA are directly obtained from original paper since we use the same data split and fine-tuning strategy as theirs. The other comparison results are obtained by re-implementing corresponding methods with their released code with the same network backbone and fine-tuning strategy as ours. We also compare with models fine-tuned with random initialization and with weights pre-trained on ImageNet data, denoted as "Random init" and "ImageNet init" respectively. We use the area under the ROC curve (AUC) on NIH ChestX-ray, CheXpert, and RSNA datasets and accuracy (ACC) on COVIDx dataset as the evaluation metric following [22,28].

Table 1 shows the results on four downstream datasets for chest X-ray classification. We can see that, compared to "Random init" and "ImageNet init" models, pre-training on medical datasets significantly improve the fine-tuning performance on all the datasets. This shows the importance of medical visual representation learning. Compared to MAE model that only uses image data for pre-training, the other methods leveraging cross-modal image-text pre-training obtain higher performance, demonstrating the great benefits of detailed description in text reports. Our CMITM model generally outperforms methods that use

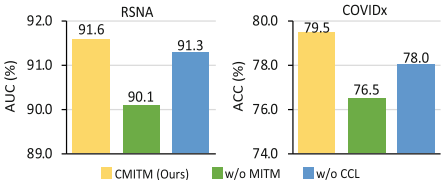


**Fig. 2.** Effects of two SSL components in our method (MITM and CCL).

**Table 2.** Ablation on removing either design in our framework.

| Methods | RSNA | COVIDx |
|---|---|---|
| CMITM (ours) | **91.6** | **79.5** |
| - cascaded training | 90.8 | 76.2 |
| - image masking | 90.9 | 77.0 |
| - representation decoder | 91.2 | 79.0 |

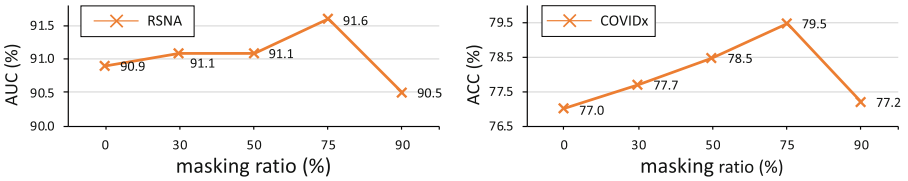


**Fig. 3.** Effect of masking ratio for cross-modal contrastive learning in our method.

either masked autoencoding or contrastive learning alone, showing the effectiveness of combining the two SSL techniques. It can be observed that our method shows the most obvious improvements in scenarios with limited data. When fine-tuning with 1% labeled data, our method outperforms the current best-performing method MRM by 1.0% on NIH ChestX-ray dataset and 1.5% on COVIDx dataset. On CheXpert dataset, MRM model shows 0.2% performance gain when increasing labeled data from 1% to 100%, but with our method, fine-tuning on 1% labeled data already outperforms MRM model fine-tuned on 100% labeled data. These results may indicate that masked autoencoding and contrastive learning benefit each other more in data-scare scenarios.

**Ablation Study.** We perform ablation analysis on RSNA and COVIDx datasets with 1% labeled data to investigate the effect of each component in the proposed method. In Fig. 2, we can see that removing either the masked image-text modeling (MITM) or cross-modal contrastive learning (CCL) in our method lead to a decrease in fine-tuning results. This again reflects the complementary role of the two SSL components. In Table 2, we show that the designs of cascaded training strategy, image masking in the contrastive learning, and representation decoder play important role to the performance of our method. Notably, results significantly decrease if not using cascaded training, that means directly using reconstruction loss and contrastive loss for joint training bring negative effects. These ablation results show that combing the two SSL approaches is non-trivial and requires careful designs to make it to be effective. For masking ratio, previous work [28] has shown that the masking ratio of 75% works well in masked medical image-text reconstruction. So we directly adopt the masking ratio of 75% for the masked image-text modeling in the first pre-training stage, but we analyze how the performance changes with the masking ratio for the

contrastive learning in the second pre-training stage. As shown in Fig. 3, it is interesting to see that the optimal masking ratio is also 75%. This might indicate that the masking ratio should keep as the same during the cascaded training.

## 4 Conclusion

We present a novel framework for medical visual representation learning by integrating the strengths of both cross-modal contrastive learning and masked image-text modeling. With careful designs, the effectiveness of our method is demonstrated on four downstream classification datasets, consistently improving data efficiency under data-scarce scenarios. This shows the complementary benefits of the two SSL techniques in medical visual representation learning. One limitation of the work is that the pre-training model is evaluated solely on classification tasks. A compelling extension of this work would be to conduct further evaluation on a broader spectrum of downstream tasks, including organ segmentation, lesion detection, and image retrieval, thereby providing a more comprehensive evaluation of our model's capabilities.

**Acknowledgements.** This work was supported by NIH R01HL159183.

## References

1. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol. 13696, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning, pp. 1597–1607. PMLR (2020)
3. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI, Singapore, September 18–22, 2022, Proceedings, Part V, vol. 13435, pp. 679–689. Springer (2022), https://doi.org/10.1007/978-3-031-16443-9_65
4. De Fauw, J., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. **24**(9), 1342–1350 (2018)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Minneapolis, MN, USA, June 2–7, 2019, vol. 1, pp. 4171–4186 (2019), https://doi.org/10.18653/v1/n19-1423
6. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR (2021)

7. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
9. 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, pp. 9729–9738 (2020)
10. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: a multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3942–3951 (2021)
11. Huang, Z., et al.: Contrastive masked autoencoders are stronger vision learners. arXiv preprint arXiv:2207.13532 (2022)
12. Irvin, J., 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)
13. Jiang, Z., et al.: Layer grafted pre-training: bridging contrastive learning and masked image modeling for label-efficient representations. arXiv preprint arXiv:2302.14138 (2023)
14. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
15. Liao, R., et al.: Multimodal representation learning via maximization of local mutual information. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part II, vol. 12902, pp. 273–283. Springer (2021), https://doi.org/10.1007/978-3-030-87196-3_26
16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Mishra, S., et al.: A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. arXiv preprint arXiv:2210.16870 (2022)
18. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest X-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI, Singapore, September 18–22, 2022, Proceedings, Part V, vol. 13435, pp. 647–657. Springer (2022), https://doi.org/10.1007/978-3-031-16443-9_62
19. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
20. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
21. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiol. Artif. Intell. **1**(1), e180041 (2019)
22. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Adv. Neural. Inf. Process. Syst. **35**, 33536–33549 (2022)
23. Wang, L., Lin, Z.Q., Wong, A.: Covid-Net: a tailored deep convolutional neural network design for detection of Covid-19 cases from chest X-ray images. Sci. Rep. **10**(1), 1–12 (2020)

24. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471 (2017)
25. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: medical knowledge enhanced language-image pre-training. medRxiv pp. 2023–01 (2023)
26. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference, pp. 2–25. PMLR (2022)
27. Zhou, H.Y., Chen, X., Zhang, Y., Luo, R., Wang, L., Yu, Y.: Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. Nature Mach. Intell. **4**(1), 32–40 (2022)
28. Zhou, H., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023 (2023)