MedIM: Boost Medical Image Representation via Radiology Report-Guided Masking

Yutong Xie[1], Lin Gu[2,3], Tatsuya Harada[2,3], Jianpeng Zhang[4], Yong Xia[4], and Qi Wu[1(✉)]

[1] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, Australia
qi.wu01@adelaide.edu.au
[2] RIKEN AIP, Tokyo, Japan
[3] RCAST, The University of Tokyo, Tokyo, Japan
[4] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China

Abstract. Masked image modelling (MIM)-based pre-training shows promise in improving image representations with limited annotated data by randomly masking image patches and reconstructing them. However, random masking may not be suitable for medical images due to their unique pathology characteristics. This paper proposes **M**asked m**ed**ical **I**mage **M**odelling (MedIM), a novel approach, to our knowledge, the first research that masks and reconstructs discriminative areas guided by radiological reports, encouraging the network to explore the stronger semantic representations from medical images. We introduce two mutual comprehensive masking strategies, knowledge word-driven masking (KWM) and sentence-driven masking (SDM). KWM uses Medical Subject Headings (MeSH) words unique to radiology reports to identify discriminative cues mapped to MeSH words and guide the mask generation. SDM considers that reports usually have multiple sentences, each of which describes different findings, and therefore integrates sentence-level information to identify discriminative regions for mask generation. MedIM integrates both strategies by simultaneously restoring the images masked by KWM and SDM for a more robust and representative medical visual representation. Our extensive experiments on various downstream tasks covering multi-label/class image classification, medical image segmentation, and medical image-text analysis, demonstrate that MedIM with report-guided masking achieves competitive performance. Our method substantially outperforms ImageNet pre-training, MIM-based pre-training, and medical image-report pre-training counterparts. Codes are available at https://github.com/YtongXie/MedIM.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43907-0_2.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 13–23, 2023.
https://doi.org/10.1007/978-3-031-43907-0_2

1 Introduction

Accurate medical representation is crucial for clinical decision-making. Deep learning has shown promising results in medical image analysis, but the accuracy of these models heavily relies on the quality and quantity of data and annotations [21]. Masked image modelling (MIM)-based pre-training approach [3,8,23] such as masked autoencoders (MAE) [8] has shown prospects in improving the image representation under limited annotated data. MIM masks a set of image patches before inputting them into a network and then reconstructs these masked patches by aggregating information from the surrounding context. This ability to aggregate contextual information is essential for vision tasks and understanding medical image analysis [24]. Recently, MIM has witnessed much success in medical domain [4–6,11,20,24] such as chest X-ray and CT image analysis.

While the random masking strategy is commonly used in current MIM-based works, randomly selecting a percentage of patches to mask. We argue that such a strategy may not be the most suitable approach for medical images due to the domain particularity. Medical images commonly present relatively fixed anatomical structures, while subtle variations between individuals, such as sporadic lesions that alter the texture and morphology of surrounding tissues or organs, may exist. These pathology characteristics may be minute and challenging to perceive visually but are indispensable for early screening and clinical diagnosis. Representation learning should capture these desired target representations to improve downstream diagnosis models' reliability, interpretability, and generalizability. Random masking is less likely to deliberately focus on these important parts. We put forward a straightforward principle, *i.e., masking and reconstructing meaningful characteristics*, encouraging the network to explore stronger representations from medical images.

We advocate utilising radiological reports to locate relevant characteristics and guide mask generation. These reports are routinely produced in clinical practice by expert medical professionals such as radiologists, and can provide a valuable source of semantic knowledge at little to no additional cost [9,17]. When medical professionals read a medical image, they will focus on areas of the image that are relevant to the patient's or clinical conditions. These areas are then recorded in a report, along with relevant information such as whether they are normal or abnormal, the location and density of abnormal areas, and any other materials about the patient's condition. By incorporating reports into the medical image representation learning, the models can simulate the professionals' gaze and learn to focus on the pathology characteristics of images.

In this paper, we propose a new approach called MedIM (**M**asked m**ed**ical **I**mage **M**odelling). MedIM aligns semantic correspondences between medical images and radiology reports and reconstructs regions masked by the guidance of learned correspondences. Especially we introduce two masking strategies: knowledge word-driven masking (KWM) and sentence-driven masking (SDM). KWM uses Medical Subject Headings (MeSH) words [14] as the domain knowledge. MeSH words provide a standardized language for medical concepts and conditions. In radiology reports, MeSH words describe imaging modalities, anatomic

locations, and pathologic findings, such as "Heart", "Pulmonary", "Vascular", and "Pneumothorax" in Fig. 1, and are important semantic components. This inspired KWM to identify regions mapped to MeSH words and generate an attention map, where the highly activated tokens indicate more discriminative cues. We utilize this attention map to selectively mask then restore the high-activated regions, stimulating the network to focus more on regions related to MeSH words during the modelling process. SDM considers multiple sentences in reports, each potentially providing independent information about different aspects of the image. It generates an attention map by identifying regions mapped to one selected sentence, enabling the network to focus on specific aspects of the image mentioned in that sentence during modelling. KWM and SDM identify different sources of discriminative cues and are therefore complementary. MedIM leverages the superiority of both strategies by simultaneously restoring images masked by KWM and SDM in each iteration. This integration creates a more challenging and comprehensive modelling task, which encourages the network to learn more robust and representative medical visual representations. Our MedIM approach is pre-trained on a large chest X-ray dataset of image-report pairs. The learned image representations are transferred to several medical image analysis downstream tasks: multi-label/class image classification and pneumothorax segmentation. Besides, our MedIM pre-trained model can be freely applied to image-text analysis downstream tasks such as image-to-text/text-to-image retrieval.

Our contributions mainly include three-fold: (1) we present a novel masking approach MedIM, which is the first work to explore the potential of radiology reports in mask generation for medical images, offering a new perspective to enhance the accuracy and interpretability of medical image representation; (2) we propose two mutual comprehensive masking strategies, KWM and SDM, that effectively identify word-level and sentence-level of discriminative cues to guide the mask generation; (3) we conduct extensive experiments on medical image and image-text downstream tasks, and the performance beats strong competitors like ImageNet pre-training, MIM-based pre-training and advanced medical image-report pre-training counterparts.

2 Approach

As shown in Fig. 1, our MedIM framework has dual encoders that map images and reports to a latent representation, a report-guided mask generation module, and a decoder that reconstructs the images from the masked representation.

2.1 Image and Text Encoders

Image Encoder. We use the vision Transformer (ViT) [7] as the image encoder$\mathcal{F}(\cdot)$. For an input medical image $\boldsymbol{x}$, it is first reshaped into a sequence of flattened patches that are then embedded and fed into stacked Transformer layers to obtain the encoded representations of visual tokens $\boldsymbol{E}_{img} = \mathcal{F}(\boldsymbol{x}) \in \mathbb{R}^{N_{img} \times C}$, where C is the encoding dimension and N_{img} denotes the number of patches.

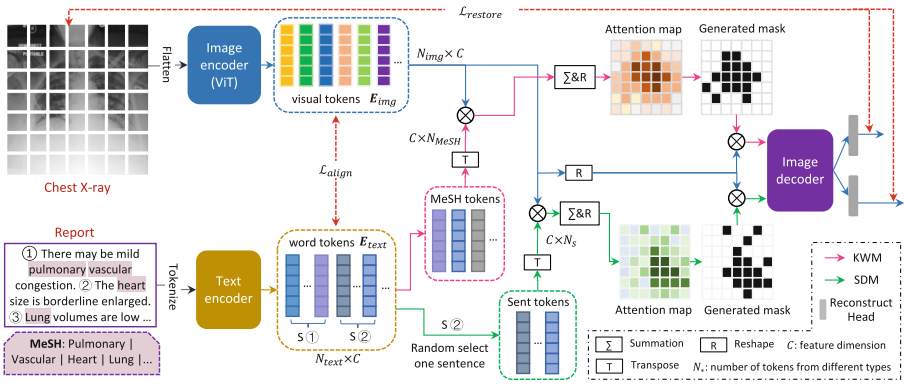


Fig. 1. Illustration of our MedIM framework. It includes dual encoders to obtain latent representations. Two report-guided masking strategies, KWM and SDM, are then introduced to generate the masked representations. The decoder is built to reconstruct the original images from the masked representation. Noted that the back regions in the generated mask will be masked.

Text Encoder. We use the BioClinicalBERT [2] model, pre-trained on the MIMIC III dataset [13], as our text encoder $\mathcal{T}(\cdot)$. We employ WordPiece [19] for tokenizing free-text medical reports. This technique is particularly useful for handling the large and diverse vocabularies that are common in the medical language. For an input medical report $\boldsymbol{r}$ with N_{text} words, the tokenizer segments each word to sub-words and generates word piece embeddings as the input to the text encoder. The text encoder extracts features for word pieces, which are aggregated to generate the word representations $\boldsymbol{E}_{text} = \mathcal{T}(\boldsymbol{r}) \in \mathbb{R}^{N_{text} \times C}$.

2.2 Report-Guided Mask Generation

We introduce two radiology report-guided masking strategies, *i.e.,* KWM and SDM, identifying different cues to guide the mask generation.

Knowledge Word-Driven Masking (KWM). MeSH words shown in Fig. 1 are important for accurately describing medical images, as they provide a standardized vocabulary to describe the anatomical structures and pathologies observed in the images. Hence the KWM is proposed to focus on the MeSH word tokens during mask generation. Given a report $\boldsymbol{r}$ and its text representations $\boldsymbol{E}_{text}$, we first match MeSH words in the report based on the MeSH Table [14] and extract the representations of MeSH word tokens, formally as

$$\boldsymbol{E}_{\text{MeSH}} = \left\{ \boldsymbol{E}_{text}^{j}, \boldsymbol{r}^{j} \in \text{MeSH}, \text{j} \in \{1, ..., \text{N}_{\text{text}}\} \right\} \in \mathbb{R}^{N_{\text{MeSH}} \times C}, \tag{1}$$

where N_{MeSH} represents the number of MeSH words in the report $\boldsymbol{r}$. Then, we compute an attention map $\mathbb{C}_{\text{MeSH}}$ to identify image regions mapped to MeSH

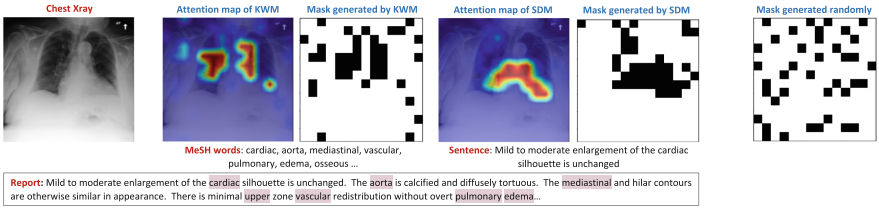


Fig. 2. A image-report pair and the corresponding attention map and mask generated by KWM and SDM. The black regions in the generated mask will be masked.

words as follows

$$\mathbb{C}_{\text{MeSH}} = R(\sum \text{softmax}(\boldsymbol{E}_{img} \cdot \boldsymbol{E}_{\text{MeSH}}^{T})) \in \mathbb{R}^{H \times W}, \tag{2}$$

where $H = W = \sqrt{N_{img}}$, T and R represent the transpose and reshape functions, and the softmax function normalizes the elements along the image dimension to find the focused region matched to each MeSH word. The summation operation $\sum$ performs on the text dimension to aggregate the attentions related to all MeSH words.

Subsequently, the high-activated masking is presented to remove the discovered attention regions. Here, we define a corresponding binary mask $\boldsymbol{m} \in \{0,1\}^{H \times W}$ formulated as $\boldsymbol{m}^{(i,j)} = \mathbb{I}(\mathbb{C}_{\text{MeSH}}^{(i,j)} \leqslant \mathbb{C}_{\text{MeSH}}^{[\gamma * N_{img}]})$. Here $\mathbb{C}_{\text{MeSH}}^{[\gamma * N_{img}]}$ refers to the $(\gamma * N_{img})$-th largest activation in $\mathbb{C}_{\text{MeSH}}$, andγ is the masking ratio that determines how many activations would be suppressed. With this binary mask, we can compute the masked representations produced by KWM as

$$\mathbb{M}(\mathbb{C}_{\text{MeSH}}; \lambda)^{kwm} = \{\boldsymbol{z}^{(i,j)} | \boldsymbol{m}^{(i,j)} \cdot R(\boldsymbol{E}_{img})^{(i,j)} + (1 - \boldsymbol{m}^{(i,j)}) \cdot \texttt{[MASK]}\}_{i=1}^{H}{}_{j=1}^{W}, \tag{3}$$

where `[MASK]` is a masked placeholder.

Sentence-Driven Masking (SDM). Medical reports often contain multiple sentences that describe different findings related to the image, which inspires SDM to introduce sentence-level information during mask generation. For the report $\boldsymbol{r}$, we randomly select a sentence $\boldsymbol{s}$ and extract its representations as

$$\boldsymbol{E}_s = \left\{ \boldsymbol{E}_{text}^{j}, \boldsymbol{r}^{j} \in s, j \in \{1, ..., N_{text}\} \right\} \in \mathbb{R}^{N_s \times C} \tag{4}$$

where N_s represents the length of $\boldsymbol{s}$. Then, an attention map $\mathbb{C}_s$ can be computed to identify regions mapped to this sentence as

$$\mathbb{C}_s = R(\sum \text{softmax}(\boldsymbol{E}_{img} \cdot \boldsymbol{E}_s^{T})) \in \mathbb{R}^{H \times W}, \tag{5}$$

After that, the high-activated masking is performed based on $\mathbb{C}_s$ to compute the masked representations $\mathbb{M}(\mathbb{C}_s; \lambda)^{sdm}$.

We also select an image-report pair and visualize the corresponding attention map and generated mask procured by KWM and SDM in Fig. 2 to show the superiority of our masking strategies.

2.3 Decoder for Reconstruction

Both masked representations $\mathbb{M}(\mathbb{C}_{\text{MeSH}};\lambda)^{kwm}$ and $\mathbb{M}(\mathbb{C}_s;\lambda)^{sdm}$ are mapped to the decoder $\mathcal{D}(\cdot)$ that includes four conv-bn-relu-upsample blocks. We design two independent reconstruction heads to respectively accept the decoded features $\mathcal{D}(\mathbb{M}(\mathbb{C}_{\text{MeSH}};\lambda)^{kwm})$ and $\mathcal{D}(\mathbb{M}(\mathbb{C}_s;\lambda)^{sdm})$ and generate the final reconstruction results $\boldsymbol{y}^{kwm}$ and $\boldsymbol{y}^{sdm}$.

2.4 Objective Function

MedIM creates a more challenging reconstruction objective by removing then restoring the most discriminative regions guided by radiological reports. We optimize this reconstruction learning process with the mean square error (MSE) loss function, expressed as

$$\mathcal{L}_{restore} = \left\|\boldsymbol{y}^{kwm}, \boldsymbol{x}\right\|^2 + \left\|\boldsymbol{y}^{sdm}, \boldsymbol{x}\right\|^2 \tag{6}$$

MedIM also combines the cross-modal alignment constraint, which aligns medical images' visual and semantic aspects with their corresponding radiological reports, benefiting in better identifying the reported-guided discriminative regions during mask generation. We follow the work [17] and compute the objective alignment function $\mathcal{L}_{align}$ by exploiting the fine-grained correspondences between images and reports. The final objective of our MedIM is the combination of reconstruction and alignment objectives as $\mathcal{L}_{\text{MedIM}} = \alpha\mathcal{L}_{\text{restore}} + \mathcal{L}_{align}$, where α is a weight factor to balance both objectives.

2.5 Downstream Transfer Learning

After pre-training, we can transfer the weight parameters of the MedIM to various downstream tasks. For the classification task, we use the commonly used Linear probing, *i.e.*, freezing the pre-trained image encoder and solely training a randomly initialized linear classification head. For the segmentation task, the encoder and decoder are first initialized with the MedIM pre-trained weights, and a downstream-specific head is added to the network. The network is then fine-tuned end-to-end. For the retrieval task, we take an image or report as an input query and retrieve target reports or images by computing the similarity between the query and all candidates using the learned image and text encoders.

3 Experiments and Results

3.1 Experimental Details

Pre-training Setup. We use the MIMIC-CXR-JPG dataset [12] to pre-train our MedIM framework. Following [17], we only include frontal-view chest images from the dataset and extract the impression and finding sections from radiological reports. As a result, over 210,000 radiograph-report pairs are available. We

manually split 80% of pairs for pre-training and 20% of pairs used for downstream to validate in-domain transfer learning. We set the input size to 224×224 adopt the AdamW optimizer [16] with a cosine decaying learning rate [15], a momentum of 0.9, and a weight decay of 0.05. We set the initial learning rate to 0.00002, batch size to 144, and maximum epochs to 50. Through the ablation study, we empirically set the mask ratio to 50% and loss weight α to 10.

Downstream Setup. We validate the transferability of learned MedIM representations on four X-ray-based downstream tasks: (1) multi-label classification on CheXpert [10] dataset using its official split, which contains five individual binary labels: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion; (2) multi-class classification on COVIDx [18] dataset with over 30k chest X-ray images, which aims to classify each radiograph into COVID-19, non-COVID pneumonia or normal, and is split into training, validation, and test set with 80%/10%/10% ratio; (3) pneumothorax segmentation on SIIM-ACR Pneumothorax Segmentation dataset [1] with over 12k chest radiographs, which is split into training, validation, and test set with 70%/15%/15% ratio; and (4) image-text/report-text retrieval on the MIMIC-CXR validation dataset. We use the Dice coefficient score (Dice) to measure the segmentation performance, use the mean area under the receiver operator curve (mAUC) to measure the multi-label classification performance, and use the accuracy to measure the multi-class classification performance. We use the recall of the corresponding image/report that appears in the top-k ranked images/reports (denoted by R@k) to measure the retrieval performance [9]. Each downstream experiment is conducted three times and the average performance is reported. More details are in the Appendix.

3.2 Comparisons with Different Pre-training Methods

We compare the downstream performance of our MedIM pre-training with five pre-training methods in Table 1 and Table 2. Our MedIM achieves state-of-the-art results on all downstream datasets, outperforming ImageNet pre-training [7], MIM-based pre-training MAE [8] and three medical image-report pre-training approaches, GLoRIA [9], MRM [22] and MGCA [17], under different labelling ratios. The superior performance corroborates the effectiveness of our report-guided masking pre-training strategy over other pre-training strategies in learning discriminative information. Besides, our MedIM achieves 88.91% when using only 1% downstream labelled data on CheXpert, better than other competitors with 100% labelled data. These convincing results have demonstrated the enormous potential of MedIM for annotation-limited medical image tasks.

3.3 Discussions

Ablation Study. Ablation studies are performed over each component of MedIM, including knowledge word-driven masking (KWM) and Sentence-driven masking (SDM), as listed in Table 3. We sequentially add each component to

Table 1. Classification and segmentation results of different pre-training methods on three downstream test sets under different ratios of available labelled data. All methods were evaluated with the ViT-B/16 backbone. * denotes our implementation of on same pre-training dataset and backbone due to the lack of available pre-trained weights.

Methods	CheXpert			COVIDx			SIIM	
	1%	10%	100%	1%	10%	100%	10%	100%
Random Init	68.11	71.17	71.91	67.01	79.68	82.71	19.13	60.97
ImageNet [7]	73.52	80.38	81.84	71.56	84.28	89.74	55.06	76.02
MAE* [8]	82.36	85.22	86.69	73.31	87.67	91.79	57.68	77.16
GLoRIA* [9]	86.50	87.53	88.24	75.79	88.68	92.11	57.67	77.23
MRM [22]	88.50	88.50	88.70	76.11	88.92	92.21	61.21	79.45
MGCA* [17]	88.11	88.29	88.88	76.29	89.04	92.47	60.64	79.31
MedIM	**88.91**	**89.25**	**89.65**	**77.22**	**90.34**	**93.57**	**63.50**	**81.32**

the vanilla baseline, $\mathcal{L}_{align}$ only, thus the downstream performance is gradually improved in Table 3. First, by reconstructing the masked representations produced by KWM, the total performance of three tasks is increased by 3.28 points. This indicates that using MeSH words as knowledge to guide the mask generation can improve the model representations and generalization. Equipped with KWM and SDM, our MedIM can surpass the baseline model by a total of 5.12 points on three tasks, suggesting the superiority of adding the SDM strategy and integrating these two masking strategies.

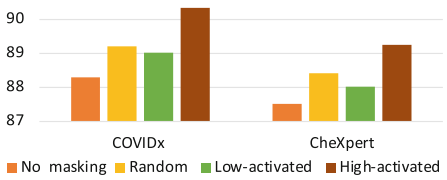
Masking Strategies. To demonstrate the effectiveness of the High-activated masking strategy, we compare it with three counterparts, No masking, Random masking, and Low-activated masking. Here No masking means that the recon-

Table 2. Image-to-text (I2T) and text-to-image (T2I) retrieval results on the MIMIC-CXR test set.

Methods	T2I			I2T		
	R@1	R@5	R@10	R@1	R@5	R@10
MGCA [17]	5.74	22.91	31.90	6.22	23.61	32.51
MedIM	**7.67**	**23.96**	**33.55**	**8.70**	**24.63**	**34.27**

Table 3. Ablation study of different components in MedIM.

Different components			Tasks		
$\mathcal{L}_{align}$	KWM	SDM	COVIDx	CheXpert	SIIM
✓	×	×	89.04	88.29	60.64
✓	✓	×	89.85	88.86	62.54
✓	✓	✓	**90.34**	**89.25**	**63.50**



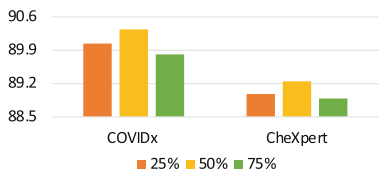


Fig. 3. Left: Results when using different masking strategies. **Right**: Results when using different masking ratios.

struction is performed based on the complete image encoder representations instead of the masked one. Low-activated masking refers to masking the tokens exhibiting a low response in both KWM and SDM strategies. The comparison on the left side of Fig. 3 reveals that all masking strategies are more effective in improving the accuracy than No masking. Benefiting from mining more discriminative information, our High-activated masking performs better than the Random and Low-activated masking. Besides, we also compare different masking ratios, varying from 25% to 75%, on the right side of Fig. 3.

4 Conclusion

We propose a new masking approach called MedIM that uses radiological reports to guide the mask generation of medical images during the pre-training process. We introduce two masking strategies KWM and SDM, which effectively identify different sources of discriminative cues to generate masked inputs. MedIM is pre-trained on a large dataset of image-report pairs to restore the masked regions, and the learned image representations are transferred to three medical image analysis tasks and image-text/report-text retrieval tasks. The results demonstrate that MedIM outperforms strong pre-training competitors and the random masking method. In the future, we will extend our MedIM to handle other modalities, *e.g.,* 3D medical image analysis.

Acknowledgments. Dr. Lin Gu was supported by JST Moonshot R&D Grant Number JPMJMS2011, Japan. Prof. Yong Xia was supported in part by the Key Research and Development Program of Shaanxi Province, China, under Grant 2022GY-084, in part by the National Natural Science Foundation of China under Grants 62171377, and in part by the National Key R&D Program of China under Grant 2022YFC2009903/2022YFC2009900.

References

1. Siim-acr pneumothorax segmentation. Society for Imaging Informatics in Medicine (2019)
2. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: BERT pre-training of image transformers. In: International Conference on Learning Representations (ICLR) (2022)
4. Cai, Z., Lin, L., He, H., Tang, X.: Uni4Eye: unified 2D and 3D self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13438, pp. 88–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16452-1_9

5. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3D medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1970–1980 (2023)
6. Chen, Z., et al.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13435, pp. 679–689. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_65
7. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
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. 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)
10. 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)
11. Jiang, J., Tyagi, N., Tringale, K., Crane, C., Veeraraghavan, H.: Self-supervised 3D anatomy segmentation using self-distilled masked image transformer (smit). In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 556–566. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_53
12. Johnson, A.E., et al.: Mimic-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 1–8 (2019)
13. Johnson, A.E., et al.: Mimic-III, a freely accessible critical care database. Sci. Data **3**(1), 1–9 (2016)
14. Lipscomb, C.E.: Medical subject headings (mesh). Bull. Med. Libr. Assoc. **88**(3), 265 (2000)
15. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: ICLR (2017)
16. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in Adam (2018)
17. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Advances in Neural Information Processing Systems (2022)
18. 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)
19. Wu, Y., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
20. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3588–3600 (2023)
21. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: UniMISS: universal medical self-supervised learning via breaking dimensionality barrier. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13681, pp. 558–575. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19803-8_33

22. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: International Conference on Learning Representations (ICLR) (2023)
23. Zhou, J., et al.: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations (ICLR) (2022)
24. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image analysis. arXiv preprint arXiv:2203.05573 (2022)