RGMIM: Region-Guided Masked Image Modeling for COVID-19 Detection

Guang Li\(^a\), Ren Togo\(^b\), Takahiro Ogawa\(^b\), Miki Haseyama\(^b\)

\(^a\)Graduate School of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-Ku, Sapporo, 060-0814, Japan
\(^b\)Faculty of Information Science and Technology, Hokkaido University, N-14, W-9, Kita-Ku, Sapporo, 060-0814, Japan

Abstract

**Background and objective:** Self-supervised learning is rapidly advancing computer-aided diagnosis in the medical field. Masked image modeling (MIM) is one of the self-supervised learning methods that masks a subset of input pixels and attempts to predict the masked pixels. Traditional MIM methods often employ a random masking strategy. In comparison to ordinary images, medical images often have a small region of interest for disease detection. Consequently, we focus on fixing the problem in this work, which is evaluated by automatic COVID-19 identification.

**Methods:** In this study, we propose a novel region-guided masked image modeling method (RGMIM) for COVID-19 detection in this paper. In our method, we devise a new masking strategy that employed lung mask information to identify valid regions to learn more useful information for COVID-19 detection. The proposed method was contrasted with five self-supervised learning techniques (MAE, SKD, Cross, BYOL, and, SimSiam). We present a quantitative evaluation of open COVID-19 CXR datasets as well as masking ratio hyperparameter studies.

**Results:** When using the entire training set, RGMIM outperformed other comparable methods, achieving 0.962 detection accuracy. Specifically, RGMIM significantly improved COVID-19 detection in small data volumes, such as 5% and 10% of the training set (846 and 1,693 images) compared to other methods, and achieved 0.957 detection accuracy even when only 50% of the training set was used.

**Conclusions:** RGMIM can mask more valid lung-related regions, facilitating the learning of discriminative representations and the subsequent high-accuracy COVID-19 detection. RGMIM outperforms other state-of-the-art self-supervised learning methods in experiments, particularly when limited training data is used. With the help of the Segment Anything Model, RGMIM also has the potential to be applied to different modalities of medical images.

**Keywords:** Self-supervised learning, Masked image modeling, Segment anything model, COVID-19 detection, Chest X-ray images.

1. Introduction

Self-supervised learning is regarded as the most essential part of artificial intelligence and has made great progress in the medical field \cite{1, 2, 3}. Early self-supervised learning techniques usually learn representations through a simple prior task that applies transformations such as rotate \cite{4} and jigsaw puzzle \cite{5} to the input image and make the model predict the properties of the transformation from the transformed image. Afterward, contrastive learning methods have been demonstrated to be efficient on different tasks \cite{6}. These approaches aim to make the distances between various view representations enhanced from the same image (i.e., positive sample pairs) closer while making the distances between view representations enhanced from different images (i.e., negative sample pairs) farther \cite{7}. Nevertheless, as these techniques depend on data augmentation, the hyperparameters and data augmentation strategies need to be changed frequently \cite{8}.

As one of the hot research topics of self-supervised learning, masked image modeling (MIM) has recently attracted more and more attention \cite{9, 10, 11}. In the discipline of natural language processing, masked language modeling (MLM) is a task that masks a portion of the input pixels and tries attempts to predict the masked pixels \cite{12}. Accordingly, recovering a large number of masked patches from a small number of unmasked patches allows the model to learn sufficient semantic information \cite{13}. In comparison to contrastive learning techniques relying on significant data augmentation, MIM is also more straightforward and reliable. As a result, it is valuable to investigate the role of MIM on real-world datasets such as medical image datasets \cite{14}. Traditional MIM techniques frequently employ a random masking strategy for ordinary images \cite{10}. However, medical images often have a small region of interest for disease detection compared to ordinary images \cite{15}. For instance, when applied to COVID-19 detection, the regions out-
side the lung may not have the information needed to make a judgment, which could prevent the random masking strategy from learning enough information.

COVID-19 and its variants are rapidly spreading around the world, affecting the health and lives of billions of people [16]. Because of the rising incidence of COVID-19 infections, many healthcare facilities are experiencing difficulties, including a severe lack of medical resources and the self-infection of many healthcare professionals [17]. Early detection of COVID-19 is crucial to stop the disease from spreading and lessen the strain on medical institutions because COVID-19 is very changeable and contagious [18]. According to the reports of the world health organization, reverse transcription polymerase chain reaction (RT-PCR) has been recognized as the gold standard for COVID-19 detection. However, some references indicate that RT-PCR is time-consuming and has a high false-negative rate [19]. Although pneumonia is seen in a substantial number of patients with confirmed COVID-19, the radiological evaluation may be useful for quick identification [20]. As a result, computer-aided diagnosis based on chest X-ray (CXR) images has the potential for the fast screening of COVID-19 during patient triage, determining the priority of patient care to support overwhelmed medical facilities in a pandemic situation.

In this study, a brand-new region-guided masked image modeling (RGMIM) technique for COVID-19 detection is proposed. In our method, we devise a new masking strategy that uses lung mask information to locate valid regions to learn more useful COVID-19 detection information. As depicted in Fig. 1, the proposed region-guided masking strategy can resolve the shortcomings of random masking in CXR images. RGMIM obtained 0.962 detection accuracy on an open COVID-19 CXR dataset according to experimental results. When limited training data was used, RGMIM drastically outperformed other state-of-the-art (SOTA) self-supervised learning methods on an open COVID-19 CXR dataset, especially when using limited training data.

Our main contributions are listed as follows:

- We propose a novel RGMIM for COVID-19 detection that uses lung mask information to locate valid regions to learn more useful COVID-19 detection information.
- RGMIM significantly outperforms other SOTA self-supervised learning methods on an open COVID-19 CXR dataset, especially when using limited training data.
- We explore the possibility of applying SAM to RGMIM and demonstrate the potential of RGMIM for different modalities of medical images.

The next of the paper is organized as follows. First, we depict the details of the proposed method in Section 2. Then, we show the experimental settings and results in Section 3. Finally, we conduct discussions and conclusions in Sections 4 and 5, respectively.

2. Methods

RGMIM is a pretraining task that uses Vision Transformer (ViT) to compare the original patches and the reconstructed masked patches [22]. The network structure contains a ViT encoder and a ViT decoder. The ViT encoder transforms the unmasked CXR patches into a latent representation. From the latent representation, the ViT decoder can reconstruct masked CXR patches. An overview of RGMIM is shown in Fig. The left indicates the pipeline and the right show the structure of the ViT encoder. RGMIM consists of five steps: (1) splitting the input CXR image into patches, (2) matching the patches with the lung mask image, (3) masking some valid patches, (4) embedding the unmasked patches, and (5) compare the reconstructed patches and masked patches. In the following subsections, we go over the region-guided masking strategy, ViT encoder and decoder, pretraining process, fine-tuning, and COVID-19 detection in detail.
2.1. Region-Guided Masking Strategy

First, given an input CXR image $X \in \mathbb{R}^{H \times W \times C}$, we divide it into $n$ patches $X_N \in \mathbb{R}^{n \times (T^2 \times C)}$, where $(H, W)$ denotes the size of the original CXR image, $C$ denotes the number of channels, $(T, T)$ denotes the size of splitted patches, and $n = HW/T^2$ is the number of patches. For ordinary images, traditional MIM methods frequently employ a random masking strategy. However, COVID-19 detection is only related to lung characteristics, and information from areas other than the lungs is meaningless. As shown on the left of Fig. 1, if we use a random masking strategy for CXR images, we may mask too many ineffective areas (e.g., areas outside the lungs), and thus the self-supervised learning process cannot learn sufficient representations.

To address the aforementioned issue, we developed a new masking strategy that employs lung mask information to locate valid regions to learn more useful information for COVID-19 detection. For example, for each input CXR image $X$, there is a corresponding lung mask image that represents the lung region. And we match the split patches $X_N$ with the lung mask image. We define split patches as valid CXR patches if they overlap with the lung mask image. We then randomly mask $m = n \times \sigma$ valid CXR patches $X_M$ with the masking ratio $\sigma$ and obtain the remaining $u = n \times (1 - \sigma)$ unmasked patches $X_U$. The unmasked patches $X_U$ served as the effective input sequence of the ViT encoder. With the proposed region-guided masking strategy, we can mask more valid lung-related regions, facilitating the learning of discriminative representations and the subsequent high-accuracy COVID-19 detection.

2.2. ViT Encoder and Decoder

Following that, because all layers of the ViT encoder use a constant latent vector size $D$, we flatten the $u$ unmasked patches and transform them to $D$ dimensions using a trainable linear projection, as shown bellow:

$$z_0 = [X_1^U E; X_2^U E; \cdots; X_u^U E] + E_{pos}.$$  \hspace{1cm} (1)

where $E \in \mathbb{R}^{(T^2 \times C) \times D}$ denotes the patch embeddings and $E_{pos} \in \mathbb{R}^{(u \times D)}$ denotes the position embeddings. Position embeddings $E_{pos}$ are appended to the patch embeddings $E$ to preserve positional information. We employ learnable one-dimensional position embeddings, and the got sequence of embedding vectors $z_0$ is input to the ViT encoder. The ViT encoder has a total number of $L$ blocks. Each block is made up of a multi-head self-attention (MSA) module, as well as a multilayer perceptron (MLP), layer normalization (LN), and residual connections. Each layer of the MLP has a Gaussian error linear unit activation function. The global average pooling (GAP) is used to compute the final output embeddings $y$ of unmasked patches $X_U$. The embedding process is defined as follows:

$$\hat{z}_l = MSA(LN(z_{l-1})), \quad l = 1...L$$ \hspace{1cm} (2)

$$z_l = MLP(LN(\hat{z}_l)), \quad l = 1...L$$ \hspace{1cm} (3)

$$y = GAP(LN(z_L)).$$ \hspace{1cm} (4)

The ViT decoder is an inverse version of the ViT encoder. The inputs of the ViT decoder are final output embeddings $y$ and masked tokens. The ViT decoder is only used in the pretraining process and can be designed differently from the ViT encoder.
The representations learned by the MIM are also well when using a lightweight ViT decoder with fewer layers, and it can significantly decrease the network parameters and pretraining time [10].

2.3. Pretraining Process

Finally, by predicting the pixel values of masked patches, and RGMIM reconstructs the original patches $X_M$. Each element in the ViT decoder output is a vector of pixel values representing a patch. Following the ViT decoder, a linear projection layer with the number of output channels equal to the number of pixel values in a patch is added. And the ViT decoder’s outputs are reshaped to generate reconstructed patches $Y_M$. We calculate the $L_2$ error between the reconstructed patches $Y_M$ and original patches $X_M$ in the pixel-level as follows:

$$L = \frac{1}{m \times T^2 \times C} \| Y_M - X_M \|_2^2,$$

(5)

where $m \times T^2 \times C$ denotes the number of elements in original patches. As a self-supervised pretraining process, the ViT encoder can learn discriminative representations from non-labeled CXR images.

2.4. Fine-Tuning and COVID-19 Detection

The ViT decoder is discarded following the RGMIM pretraining process, and the pretrained ViT encoder with a linear classification layer is used to fine-tune labeled CXR images using the cross-entropy loss. Given unknown CXR images, the fine-tuned ViT encoder with a linear classification layer can predict the categories of the input CXR images and perform COVID-19 detection during the test phase. Since the ViT encoder has learned discriminative representations from CXR images, the fine-tuning process can be important shortened and still achieve great detection accuracy with limited labeled CXR images.

### Table 1: Hyperparameters of the proposed method.

| Hyperparameter       | Value |
|----------------------|-------|
| Pretraining epoch    | 40    |
| Fine-tuning epoch    | 30    |
| Batch size           | 256   |
| Base Learning rate   | 1.5e-4|
| Weight decay         | 0.05  |
| Momentum $\beta_1$   | 0.9   |
| Momentum $\beta_2$   | 0.95  |
| Patch size $T$       | 16    |
| Masking ration $\sigma$ | 75%  |

3. Experiments

In this section, we perform extensive experiments and hyperparameter studies to validate RGMIM’s effectiveness. The dataset and settings are depicted in subsection 3.1. The findings and analysis are presented in subsection 3.2.

3.1. Dataset and Settings

The dataset used in our study is an open COVID-19 CXR image dataset (COVID) [27]. CXR images were gathered from a variety of publicly available datasets, online sources, and published articles. Each CXR image has a corresponding lung mask image. The lung mask images are confirmed by expert radiologists as ground truth. The COVID dataset has 21,165 CXR images. More information about the COVID dataset can be discovered on the official website[1]. Figure 3 shows examples of CXR images in four categories. We randomly choose 80% of the CXR images in each category as the training set (i.e., 16,933

---

[1] https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database
Table 2: COVID-19 detection accuracy of different fine-tuning data volumes.

| Method   | Architecture | 1%  | 5%  | 10% | 50% | 100% |
|----------|--------------|-----|-----|-----|-----|------|
| RGMIM    | ViT-Base     | 0.771 | 0.893 | 0.919 | 0.957 | 0.962 |
| MAE [10] | ViT-Base     | 0.754 | 0.875 | 0.903 | 0.948 | 0.956 |
| SKD [23] | ResNet-50    | 0.742 | 0.812 | 0.896 | 0.947 | 0.957 |
| Cross [24] | ResNet-50    | 0.747 | 0.795 | 0.817 | 0.934 | 0.953 |
| BYOL [25] | ResNet-50    | 0.683 | 0.754 | 0.790 | 0.933 | 0.954 |
| SimSiam [26] | ResNet-50    | 0.623 | 0.700 | 0.781 | 0.929 | 0.949 |
| Transfer | ViT-Base     | 0.689 | 0.861 | 0.893 | 0.940 | 0.953 |
| Transfer | ResNet-50    | 0.539 | 0.619 | 0.665 | 0.913 | 0.936 |
| From Scratch | ViT-Base   | 0.413 | 0.580 | 0.645 | 0.810 | 0.848 |
| From Scratch | ResNet-50 | 0.284 | 0.496 | 0.532 | 0.619 | 0.774 |

Table 3: Hyperparameter studies on masking ratio.

| Method   | Architecture | Masking ratio | Accuracy |
|----------|--------------|---------------|----------|
| RGMIM    | ViT-Base     | 0.15          | 0.949    |
|          |              | 0.30          | 0.953    |
|          |              | 0.45          | 0.955    |
|          |              | 0.60          | 0.958    |
|          |              | 0.75          | 0.962    |
|          |              | 0.90          | 0.957    |
| MAE [10] | ViT-Base     | 0.15          | 0.936    |
|          |              | 0.30          | 0.942    |
|          |              | 0.45          | 0.947    |
|          |              | 0.60          | 0.951    |
|          |              | 0.75          | 0.956    |
|          |              | 0.90          | 0.953    |

For SKD, Cross, BYOL, and SimSiam, we employed ResNet-50 as the backbone network. All of these contrastive learning methods make use of standard data augmentations (e.g., flipping, resizing, cropping, Gaussian blur). For Transfer and From Scratch, we used both ViT-Base and ResNet-50. To test RGMIM’s effectiveness in different training data volumes, we also randomly selected data at 1%, 5%, 10%, and 50% of the total dataset size for fine-tuning, with the ratio selected in each category being the same. Because the information density of the image is lower than language, different from the MLM employing a low masking ratio (e.g., 0.15), we set the masking ratio σ to 0.75 for CXR images. We also studied the masking ratio for RGMIM and MAE using hyperparameters. We used a four-class accuracy as the evaluation metric.

3.2. Comparing to SOTA Self-Supervised Learning Methods and Baseline Methods

The COVID-19 detection accuracy in different training data volumes is displayed in Table 2. Table 2 shows that RGMIM outperformed other comparative methods, achieving 0.962 detection accuracy when using the entire training set. Specifically, compared to other methods, RGMIM significantly im-
Table 4: Performance comparison with the existing methods.

| Method                     | Architecture           | Dataset                      | Accuracy                  |
|----------------------------|------------------------|------------------------------|---------------------------|
| RGMIM (Ours)               | ViT-Base               | COVID-19: 3,616, Normal: 10,192, Viral Pneumonia: 1,345, Lung Opacity: 6,012 | Four-class: 0.962          |
| T. Rahman et al. [27]      | ChexNet                | COVID-19: 3,616, Normal: 8,851, Lung Opacity: 6,012 | Three-class: 0.963         |
| M. Gour et al. et al. [28] | UA-ConvNet             | COVID-19: 219, Normal: 1,341, Viral Pneumonia: 1,345 | Three-class: 0.988         |
| L. Wang et al. [29]        | COVID-Net              | COVID-19: 358, Normal: 8,066, Viral Pneumonia: 5,538 | Three-class: 0.933         |
| Ensemble models: [30]      | VGG16, ResNet152, DenseNet201 | COVID-19: 423, Normal: 1,579, Viral Pneumonia: 1,485 | Three-class: 0.962         |
| M. Togacar et al. [31]    | Stacked models: MobileNetV2, SqueezeNet, SVM | COVID-19: 295, Normal: 65, Viral Pneumonia: 98 | Three-class: 0.993         |
| R. Zhang et al. [32]      | ResNet34               | COVID-19: 189, Normal: 235, Viral Pneumonia: 63 | Three-class: 0.911         |
| S. Minaee et al. [33]     | SqueezeNet             | COVID-19: 520, Normal: 5000 | Two-class: 0.955           |
| T. Ozturk et al. [34]     | DarkCovidNet           | COVID-19: 127, Normal: 500 | Two-class: 0.981           |
| A. Waheed et al. [35]     | Auxiliary Classifier Generative Adversarial Network | COVID-19: 403, Normal: 721 | Two-class: 0.950           |
| A. Narin et al. [36]      | Inception-ResNetV2     | COVID-19: 50, Normal: 50 | Two-class: 0.980           |

proved COVID-19 detection in small data volumes, such as 5% and 10% of the training set (846 and 1,693 images), and achieved 0.957 detection accuracy even for 50% of the training set. We can see from Table 2 that the knowledge learned by RGMIM is useful for self-supervised learning on CXR images and improves representation learning performance for COVID-19 detection. Table 3 reveals the COVID-19 detection accuracy of RGMIM and MAE when employing different masking ratios. From Table 3 when the masking ratio \( \sigma = 0.75 \), the best COVID-19 detection accuracy is achieved, indicating that the proper masking ratio can improve the detection accuracy of both methods. In addition, we observe that RGMIM outperforms MAE in terms of robustness, especially when the masking ratio is relatively low, demonstrating the superiority of our proposed method in handling incomplete CXR images.

In Figure 4, we present the accuracy changes of RGMIM when using all CXR images for fine-tuning. The results indicate that RGMIM exhibits superior accuracy and faster convergence speed compared to MAE. Specifically, after only ten epochs of learning, the detection accuracy of RGMIM has already begun to converge, while MAE is still in the process of convergence. The observed difference in convergence speed between RGMIM and MAE can be attributed to the fact that RGMIM incorporates the spatial information of CXR images through the region-guided masking strategy, which helps it to focus on the most informative regions of the images during training. Consequently, RGMIM can achieve higher accuracy with fewer iterations, making it an efficient and practical solution for COVID-19 detection from CXR images.

3.3. Comparing to Existing Methods for COVID-19 Detection

Table 4 presents the performance comparison of RGMIM with existing methods for COVID-19 detection from CXR images. While the previous methods have shown relatively high detection accuracy [27, 28, 29, 30, 31, 32, 33, 34, 35, 36], they have been evaluated on relatively small CXR image datasets with only two or three classes, which could limit their generalizability in real clinical settings. In contrast, our method was evaluated on a large CXR dataset consisting of 4 classes and 3,616 COVID-19 images and demonstrated promising detection performance. Moreover, our method utilized a simple architecture, ViT-Base, without stacked or ensemble models, which offers practicality and reliability advantages.
3.4. Exploring the Potential of RGMIM for Different Modalities of Medical Images

The proposed RGMIM approach is designed to utilize the ground truth mask image to effectively locate the relevant information within CXR images. However, in the case of other medical images with different modalities, obtaining corresponding ground truth mask images may not be feasible. To address this issue, we investigate the potential of utilizing the recently proposed Segment Anything Model (SAM) [21], a state-of-the-art technique for zero-shot image segmentation in real-world scenarios.

To evaluate the performance of SAM on medical images, we conduct experiments on three different modalities: CXR, CT, and MRI images. Specifically, we apply SAM to these medical images and assess the quality of the segmentation results. The segmentation results are presented in Fig. 5, where it can be observed that SAM is capable of accurately segmenting the regions of interest for each of the different modalities. The successful application of SAM to segment medical images of different modalities demonstrates the potential for utilizing RGMIM on these images without the need for corresponding ground truth mask images. This enables the deployment of RGMIM on a wider range of medical images and facilitates the development of more effective medical image analysis techniques.

4. Discussion

Our findings reveal the effectiveness of RGMIM for COVID-19 detection from CXR images. RGMIM outperformed other SOTA methods significantly when fine-tuning with a small amount of labeled training data. Because of the vastly different infection statuses, medical resources, and data-sharing policies of COVID-19 in various countries and cities, there is a likely risk of limited labeled training data [40]. Nonetheless, RGMIM can still be added to this case for high-accuracy COVID-19 detection. Furthermore, with the help of SAM, RGMIM has the
potential to be employed in other medical images that have a small region of interest. RGMIM also has some limitations. For example, the use of additional lung mask images and a region-guided masking strategy raises the MIM computation cost slightly. Furthermore, there are some privacy protection issues in employing medical data from patients in real-world clinical situations. However, we have previously demonstrated that medical dataset distillation improves the efficiency and security of data sharing among various medical facilities, which can be applied to RGMIM for clinical use.

5. Conclusion

This paper has suggested a novel RGMIM method for COVID-19 detection. In our method, we devise a new masking strategy that employs lung mask information to identify valid regions to learn more useful information for COVID-19 detection. RGMIM achieved 0.962 detection accuracy on an open COVID-19 CXR dataset, according to experimental results. Furthermore, when limited training data was used, RGMIM outperformed other SOTA self-supervised learning methods for COVID-19 detection. RGMIM also has the potential to be applied to other medical images. As one of our future works, we will test RGMIM on different modalities of medical image datasets for any potential bias.

Ethical approval

No ethics approval is required.

Declaration of competing interest

None declared.

Acknowledgments

This research was supported in part by the MEXT Doctoral Program for Data-Related InnoVation Expert Hokkaido University (D-DRIVE-HU) Program, AMED Grant Number JP21zf0127004, and the Hokkaido University-Hitachi Collaborative Education and Research Support Program. All experiments were conducted on the Data Science Computing System of Education and Research Center for Mathematical and Data Science, Hokkaido University.

References

[1] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, D. Rueckert, Self-supervised learning for medical image analysis using image context restoration, Medical Image Analysis 58 (2019) 101539 (2019).
[2] R. Krishnan, P. Rajpurkar, E. J. Topol, Self-supervised learning in medicine and healthcare, Nature Biomedical Engineering (2022) 1–7 (2022).
[3] S. Shurrab, R. Duwairi, Self-supervised learning methods and applications in medical imaging analysis: A survey, PeerJ Computer Science 8 (2022) e1045 (2022).
[4] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, in: Proceedings of the International Conference on Learning Representations (ICLR), 2018 (2018).
[5] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 69–84 (2016).
[6] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Transactions on Knowledge and Data Engineering (2021).
[7] T. Chen, S. Komblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the International Conference on Machine Learning (ICML), 2020 (2020).
[8] G. Li, R. Togo, T. Ogawa, M. Haseyama, Tribyol: Triplet byol for self-supervised representation learning, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3458–3462 (2022).
[9] H. Bao, L. Dong, F. Wei, Betti: Bert pre-training of image transformers, in: Proceedings of the International Conference on Learning Representations (ICLR), 2022 (2022).
[10] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16000–16009 (2022).
[11] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, H. Hu, Simmim: A simple framework for masked image modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9653–9663 (2022).
[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019 (2019).
[13] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked feature prediction for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14668–14678 (2022).
[14] Z. Chen, D. Agarwal, K. Aggarwal, W. Safta, M. M. Balan, K. Brown, Masked image modeling advances 3d medical image analysis, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 1969–1979 (2023).
[15] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Molura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: Can architectures, dataset characteristics and transfer learning, IEEE Transactions on Medical Imaging 35 (5) (2016) 1285–1298 (2016).
[16] D. Wu, T. Wu, Q. Liu, Z. Yang, The sars-cov-2 outbreak: what we know, International Journal of Infectious Diseases 94 (2020) 44–48 (2020).
[17] S. Grover, D. Dua, S. Sahoo, A. Mehra, R. Nehra, S. Chakrabarti, Why all covid-19 hospitals should have mental health professionals: The importance of mental health in a worldwide crisis!, Asian Journal of Psychiatry 51 (2020) 102147 (2020).
[18] S. A. Byrnes, R. Gallagher, A. Steadman, C. Bennett, Rivera, C. Ortega, S. T. Moltry, P. Jain, B. H. Weigl, J. T. Connelly, Multiplexed and extraction-free amplification for simplified sars-cov-2 rt-pcr tests, Analytical Chemistry 93 (9) (2021) 4160–4165 (2021).
[19] A. Boreschi, R. Maroldi, Covid-19 outbreak in italy: experimental chest x-ray scoring system for quantifying and monitoring disease progression, La Radiologia Medica 125 (5) (2020) 509–513 (2020).
[20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, arXiv:2304.02643 (2023).
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations (ICLR), 2021 (2021).
[24] G. Li, R. Togo, T. Ogawa, M. Haseyama, Covid-19 detection based on self-supervised transfer learning using chest x-ray images, International Journal of Computer Assisted Radiology and Surgery (2022) 1–8 (2022).

[25] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Geshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent-a new approach to self-supervised learning, in: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 21271–21284 (2020).

[26] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15750–15758 (2021).

[27] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughai, M. S. Khan, M. E. H. Chowdhury, Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images, Computers in Biology and Medicine 132 (2021) 104319 (2021).

[28] M. Gour, S. Jain, Uncertainty-aware convolutional neural network for covid-19 x-ray images classification, Computers in Biology and Medicine 140 (2022) 105047 (2022).

[29] L. Wang, Z. Q. Lin, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, Scientific Reports 10 (1) (2020) 1–12 (2020).

[30] N. Gianchandani, A. Jaiswal, D. Singh, V. Kumar, M. Kaur, Rapid covid-19 diagnosis using ensemble deep transfer learning models from chest radiographs to exploit social mimic optimization and structured chest x-ray images using fuzzy color and stacking approaches, Computers in Biology and Medicine 121 (2020) 103805 (2020).

[31] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, P. R. Pinheiro, Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection, IEEE Access 8 (2020) 91916–91923 (2020).

[32] L. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255 (2009).

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255 (2009).

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778 (2016).

[35] J. Shuja, E. Alanazi, W. Alasley, A. Alashaikh, Covid-19 open source data sets: a comprehensive survey, Applied Intelligence 51 (2021) 1296–1325 (2021).

[36] G. Li, R. Togo, T. Ogawa, M. Haseyama, Soft-label anonymous gastric x-ray image distillation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2020, pp. 305–309 (2020).

[37] G. Li, R. Togo, T. Ogawa, M. Haseyama, Compressed gastric image generation based on soft-label dataset distillation for medical data sharing, Computer Methods and Programs in Biomedicine 227 (2022) 107189 (2022).