Hierarchical Vision Transformers for Disease Progression Detection in Chest X-Ray Images

Amarachi B. Mbakwe[1], Lyuyang Wang[2], Mehdi Moradi[2], and Ismini Lourentzou[1(✉)]

[1] Virginia Tech, Blacksburg, VA, USA
{bmamarachi,ilourentzou}@vt.edu
[2] McMaster University, Hamilton, ON, Canada
{wangl307,moradm4}@mcmaster.ca

Abstract. Chest radiography is a commonly used diagnostic imaging exam for monitoring disease progression and treatment effectiveness. While machine learning has made significant strides in tasks such as image segmentation, disease diagnosis, and automatic report generation, more intricate tasks such as disease progression monitoring remain fairly underexplored. This task presents a formidable challenge because of the complex and intricate nature of disease appearances on chest X-ray images, which makes distinguishing significant changes from irrelevant variations between images challenging. Motivated by these challenges, this work proposes `CheXRelFormer`, an end-to-end siamese Transformer disease progression model that takes a pair of images as input and detects whether the patient's condition has improved, worsened, or remained unchanged. The model comprises two hierarchical Transformer encoders, a difference module that compares feature differences across images, and a final classification layer that predicts the change in the patient's condition. Experimental results demonstrate that `CheXRelFormer` outperforms previous counterparts. Code is available at https://github.com/PLAN-Lab/CheXRelFormer.

Keywords: Vision Transformers · Disease Progression · Chest X-Ray Comparison Relations · Longitudinal CXR Relationships

1 Introduction

Chest X-rays (CXRs) are frequently used for disease detection and disease progression monitoring. However, interpreting CXRs can be challenging and time-consuming, particularly in regions with a shortage of radiologists. This can lead to delayed or inaccurate diagnoses and management, potentially harming

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43904-9_66.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 685–695, 2023.
https://doi.org/10.1007/978-3-031-43904-9_66

patients. Automating the CXR interpretation process can lead to faster and more accurate diagnoses. Advances in Artificial Intelligence (AI), particularly in the field of computer vision for medical imaging, have significantly alleviated the challenges faced in radiology. The availability of large labeled collections of CXRs has been instrumental in driving progress in this area [8,10]. Both CXR disease detection and automatic report generation have witnessed substantial improvements [15,16,23]. Remarkably, AI-based methods for finding detection are now approaching the performance level of experienced radiologists [21,27]. Moreover, just as vision transformers have revolutionized various areas of computer vision [12], they have also become an integral part of automatic CXR analysis [18].

Although significant strides have been made in AI-assisted medical image segmentation and disease detection, tasks requiring intricate reasoning have received less attention. One such complex task is monitoring disease progression in a sequence of images, which is particularly critical in assessing patients with pneumonia and other CXR findings. For example, temporal lung changes serve as vital indicators of patient outcomes and are routinely mentioned in radiology reports for determining the course of treatment [20]. Prior work has investigated tracking the progression of COVID-19 pulmonary diseases and predicting outcomes [13]. Recently, CheXRelNet was proposed, which utilizes graph attention networks to capture anatomical correlations and detect changes in CXRs using both local and global anatomical information [11]. Change detection between longitudinal patient visits has also been studied in modalities beyond CXR, such as osteoarthritis in knee radiographs and retinopathy in retinal photographs [14]. Nonetheless, prior works have faced limitations in effectively attending to fine-grained relevant changes while disregarding irrelevant variations. Additionally, it is important to capture long-range spatial and temporal information to identify pertinent changes in medical images effectively.

Inspired by the success of Transformer models in remote sensing change detection tasks [3,28], we introduce `CheXRelFormer`, an end-to-end siamese disease progression model. `CheXRelFormer` takes a pair of CXR images as input, extracts visual features with a hierarchical vision Transformer module, and subsequently computes multi-level feature differences with a difference module. A self-attention mechanism allows the model to identify the most informative regions of the input images. By attending to fine-grained relevant changes, our model can accurately detect whether the patient's condition has improved, worsened, or remained unchanged. We evaluate the performance of our model on a large dataset of paired CXR images with corresponding disease progression labels. Experimental results demonstrate that our model outperforms existing state-of-the-art methods in detecting disease progression in CXR images. Our model has the potential to improve the efficiency and accuracy of CXR interpretation and thereby lead to more personalized treatment plans for patients. The contributions of our work can be summarized as follows:

(1) We propose `CheXRelFormer`, an end-to-end siamese disease progression model that can accurately detect changes in CXR image pairs by attend-

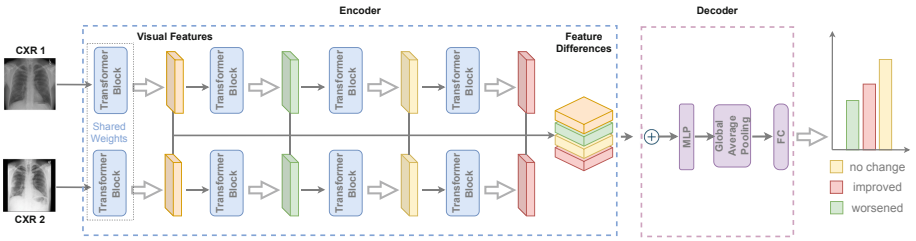

Fig. 1. `CheXRelFormer` overview. Given two CXR images $(\mathbf{X}, \mathbf{X}')$ a shared Transformer encoder extracts visual features at multiple resolutions. These feature maps are then fed into a difference module that computes visual differences across multiple scales. The difference module enhances the ability of the model to capture disease progression by effectively attending to relevant changes in the image pair. The decoder fuses information and performs the final disease progression classification task.

ing to informative regions and identifying fine-grained relevant visual differences.

(2) `CheXRelFormer` leverages hierarchical vision Transformers and a difference module to compute multi-level feature differences across CXR images, allowing the model to capture long-range spatial and temporal information.

(3) We experimentally demonstrate that `CheXRelFormer` outperforms existing state-of-the-art baselines in detecting disease progression in CXR images.

2 Methodology

Let $\mathcal{C} = \{(\mathbf{X}, \mathbf{X}')_i\}_{i=1}^{N}$ be a set of CXR image pairs, where $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{H \times W \times C}$, and H, W, and C are the height, width, and number of channels, respectively. Each image pair $(\mathbf{X}, \mathbf{X}')_i$ is associated with a set of labels $\mathcal{Y}_i = \{y_{i,m}\}_{m=1}^{M}$, where $y_{i,m} \in \{0, 1, 2\}$ indicates whether the pathology m appearing in the image pair has improved, worsened, or remained the same. The goal is to design a model that accurately predicts the disease progression labels for an unseen image pair $(\mathbf{X}, \mathbf{X}')$ and a wide range of pathologies.

To this end, we use a hierarchical Transformer [12] encoder to process each image pair. Specifically, let $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{H \times W \times C}$ be the input image pair. The encoder consists of L identical Transformer layers, each with a multi-head self-attention block followed by a position-wise feedforward network. The multi-head self-attention block contains a series of self-attention heads and is defined as

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_J)\mathbf{W}^O, \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times C}$ are the queries, keys, and values, respectively; $\mathbf{W}^O$ is a learned weight matrix, J is the number of heads, and head_j is the j-th attention head, computed as

$$\text{head}_j = \text{Attention}(\mathbf{Q}_j, \mathbf{K}_j, \mathbf{V}_j) = \text{softmax}\left(\frac{\mathbf{Q}_j \mathbf{K}_j^T}{\sqrt{d_k}}\right)\mathbf{V}_j. \tag{2}$$

Here, d_k is the dimensionality of the key and query vectors in each head, and the softmax function is applied along the rows of the matrix. The queries, keys, and values $\mathbf{Q}_j$, $\mathbf{K}_j$, and $\mathbf{V}_j$ are obtained via a set of linear projection matrices as

$$\mathbf{Q}_j = \mathbf{X}\mathbf{W}_j^Q, \quad \mathbf{K}_j = \mathbf{X}\mathbf{W}_j^K, \quad \mathbf{V}_j = \mathbf{X}\mathbf{W}_j^V, \tag{3}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are learned weight tensors that project the input embeddings onto a lower-dimensional space. Similarly, the query, key, and value matrices for the second image in the pair are computed as

$$\mathbf{Q}_j = \mathbf{X}'\mathbf{W}_j^Q, \quad \mathbf{K}_j = \mathbf{X}'\mathbf{W}_j^K, \quad \mathbf{V}_j = \mathbf{X}'\mathbf{W}_j^V, \tag{4}$$

where the weight tensors $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are shared across the two images in the pair. The output of each multi-head self-attention block for each image pair, denoted by $(\mathbf{F}, \mathbf{F}')$, is then fed into a position-wise feedforward network which consists of two linear transformations and a depth-wise convolution [4] that captures local spatial information:

$$\mathbf{F}_c = f_1(\text{Conv2D}(f_2(\mathbf{F}), \mathbf{W}_d)) \tag{5}$$

$$\mathbf{F}'_c = f_1(\text{Conv2D}(f_2(\mathbf{F}'), \mathbf{W}_d)). \tag{6}$$

Here, $\mathbf{W_d}$ is the shared depth-wise convolution weight matrix and each feed-forward layer f_1, f_2 consists of a linear transformation followed by a non-linear activation. The difference module then processes the visual features from each Transformer layer to compute multi-level feature differences as follows:

$$\mathbf{C}^l = \phi\Big(\text{Conv2D}\left([\mathbf{F}_l, \mathbf{F}'_l], \mathbf{W}_l\right)\Big). \tag{7}$$

Here, $l = 1, \ldots, L$ denotes the l-th Transformer layer, with initial inputs $(\mathbf{F}_1, \mathbf{F}'_1) = (\mathbf{F}_c, \mathbf{F}'_c)$. Furthermore, ϕ is a non-linear activation, $\mathbf{W}_l$ is a learned weight parameter that essentially represents a multi-scale trainable distance metric, and $[\cdot, \cdot]$ is the concatenation operation. By computing differences between features at different scales, the proposed model can capture local and global structures that are relevant to the disease progression task.

Each multi-scale feature difference map is then passed through a feed-forward layer that maps the input features to a common feature space

$$\mathbf{C}^l_{out} = f_{\theta_l}(\mathbf{C}^l), \ \forall l \in [1, L], \tag{8}$$

where θ_l represents the set of learnable parameters for the l-th feed-forward network. The concatenated feature tensor combines information from multiple scales and is denoted as

$$\mathbf{C}_{out} = \Big[\mathbf{C}^1_{out}, \ldots, \mathbf{C}^l_{out}, \ldots, \mathbf{C}^L_{out}\Big], \tag{9}$$

where $[\cdot]$ denotes concatenation along the channel dimension. A feed-forward network with a global average pooling step, denoted by g, creates a fused feature

representation with fixed dimensionality, which is finally passed through the final classification layer, denoted by h, to obtain the label predictions

$$\hat{y} = h\big(g(\mathbf{C}_{out})\big). \tag{10}$$

The network is trained end-to-end with a multi-label cross-entropy classification loss

$$\mathcal{L} = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{m=1}^{M} y_{i,m}\log(\sigma(\hat{y}_{i,m})) + (1 - y_{i,m})\log(1 - \sigma(\hat{y}_{i,m})), \tag{11}$$

where σ represents the sigmoid function and $\hat{y}_{i,m}, y_{i,m}$ are the model prediction and the ground truth for example $(\mathbf{X}, \mathbf{X}')_i$. An overview of the model architecture is represented in Fig. 1.

3 Experiments

3.1 Implementation Details

`CheXRelFormer` is implemented in Pytorch [19]. The encoder comprises four Transformer blocks with embedding dimensions 32, 64, 128, and 256, respectively. The number of heads on each multi-head attention block is 2, 2, 4, and 8. We train the encoder with a stochastic depth decay rule [6], with depths 3, 3, 6, and 18, for each Transformer block. To decrease the spatial dimension and reduce complexity, we perform spatial-reduction operations [22]. The position-wise feedforward network uses Gaussian Error Linear Unit (GELU) [5] activation functions. The multi-level image features extracted from the Transformer encoder are passed to the difference module. The difference module is composed of 2D convolutions, ReLU activations [2], and Batch Normalization [7]. The feed-forward layers consist of 64 neurons. The outputs from the difference module are upsampled, concatenated, and passed through a linear fusion layer followed by global average pooling. The model is trained using AdamW optimizer [17] with a learning rate of 6×10^{-5} and 16 batch size.

3.2 Dataset

We make use of the Chest ImaGenome dataset [25], which comprises 242,072 frontal MIMIC-CXRs [9] that were locally labeled using a combination of rule-based natural language processing (NLP) and CXR atlas-based bounding box detection techniques [24,26] to generate the annotations. Chest ImaGenome is represented as an anatomy-centered scene graph with 1,256 combinations of relation annotations between 29 CXR anatomical locations and their attributes. Each image is structured as one scene graph, resulting in approximately 670,000 localized comparison relations between the anatomical locations across sequential exams. In this work, we focus on the localized comparison relations data

Table 1. Dataset Statistics: pathology ID and label, number of training, validation and test CXR image pairs, and total number of CXR pairs

ID	Pathology Label	Train	Val	Test	Total Pairs
LO	Lung Opacity	5,516	912	1,579	8,007
PE	Pleural Effusion	7,450	1,089	2,165	10,704
AT	Atelectasis	77	11	26	114
EC	Enlarged Cardiac Silhouette	5,251	715	1,555	7,521
HO	Hazy Opacity/Pulmonary Edema	2,939	454	884	4,277
PX	Pneumothorax	979	132	261	1,372
CO	Consolidation	142	15	46	203
HF	Heart Failure/Fluid Overload	791	113	223	1,127
PN	Pneumonia	1,757	283	543	2,583
	Total	24,902	3,724	7,282	35,908

within Chest ImaGenome that pertains to cross-image relations for nine diseases of interest. Each comparison relation in the Chest ImaGenome dataset includes the DICOM identifiers of the two CXRs being compared, the comparison label, and the disease label name. The comparison is labeled as "no change", "improved" or "worsened", which indicates whether the patient's condition w.r.t. the disease has remained stable, improved, or worsened, respectively. The dataset contains 122, 444 unique comparisons. We use 35, 908 CXR pairs in total that pertain to the nine diseases of interest. The distribution of the data is improved (12,396), worsened (12,287) and no change (11,205). Table 1 presents high-level dataset statistics and training/validation/test splits employed in our experiments.

3.3 Baselines

To assess the performance of the proposed `CheXRelFormer` model, we conduct a comparative analysis with several baselines.

Local: This model employs a previously proposed siamese network [11] that only focuses on specific regions of the image, without considering inter-region dependencies or global information. The Local model is essentially a siamese network with a pretrained ResNet101 autoencoder trained on cropped Regions-of-Interest (RoIs), which are available in the Chest ImaGenome dataset.

Global: The Global model is a siamese network similar to the Local model but encodes global image-level information.

CheXRelNet: CheXRelNet combines global image-level information with local intra-image and inter-image information [11]. This model consists of a 2-layer graph neural network with a ResNet101 autoencoder for feature extraction.

Table 2. Quantitative comparison against the baselines

Method	LO	PE	AT	EC	HO	PX	CO	HF	PN	**All**
Local	0.41	0.37	0.41	0.29	0.37	**0.37**	0.49	0.29	0.42	0.43
Global	0.45	0.47	0.44	0.48	0.48	0.36	0.47	0.50	0.43	0.45
CheXRelNet	**0.49**	0.47	0.44	**0.49**	0.49	0.36	0.47	0.44	0.47	0.47
`CheXRelFormer`	0.48	**0.51**	**0.54**	0.40	**0.58**	0.35	**0.59**	**0.53**	**0.51**	**0.49**

Table 3. Ablation analysis of `CheXRelFormer` variants.

Method	LO	PE	AT	EC	HO	PX	CO	HF	PN	**All**
`CheXRelFormer_AbsDiff`	0.35	0.37	0.27	0.29	0.35	**0.37**	0.24	0.34	0.38	0.34
`CheXRelFormer_Local`	0.33	0.39	0.12	0.26	0.38	0.27	0.24	0.46	0.49	0.35
`CheXRelFormer`	**0.48**	**0.51**	**0.54**	0.40	**0.58**	0.35	**0.59**	**0.53**	**0.51**	**0.49**

3.4 Experimental Results

Table 2 lists the CXR change detection accuracy of all models across the nine diseases. We also report the mean weighted overall accuracy. `CheXRelFormer` outperforms baselines with a mean accuracy of 0.493 ± 0.0012 in this three-way classification task. The closest baseline is CheXRelNet with an accuracy of 0.468 ± 0.0041. Additionally, we perform a one-tailed t-test between `CheXRelFormer` and CheXRelNet, with $p = 0.00027$ indicating that `CheXRelFormer` significantly outperforms CheXRelNet in seven of the nine diseases (pleural effusion, atelectasis, pulmonary edema/hazy opacity, heart failure, pneumonia, and consolidation). Most importantly, we observe up to 12% performance gains for pathology labels with limited amounts of data, such as atelectasis and consolidation. These findings suggest that `CheXRelFormer` has the potential to be a valuable tool for detecting changes in CXR images associated with various common diseases.

3.5 Ablations on `CheXRelFormer` Architecture Components

We perform an ablation study to understand the impact of four factors, the difference module, the use of global vs. localized visual information, and the impact of multi-level features. Specifically, in Table 3, we present a comparison of in `CheXRelFormer` against `CheXRelFormer_AbsDiff`, where we replace the proposed difference module with an absolute difference component that subtracts two visual features from the last Transformer block. We also consider `CheXRelFormer_Local`, a variant that is trained on cropped anatomical RoIs instead of the entire image. Additional model variants are presented in the supplementary material. Table 4 presents the number of trainable parameters (*i.e.*, model capacity) and training time per epoch for each model variant.

Table 4. Model Capacity Comparison

Measures	CheXRelFormer_AbsDiff	CheXRelFormer_Local	CheXRelFormer
Number of Parameters (M)	28.9	41.0	41.0
Time per epoch (minutes)	148.8	56.4	171

An interesting observation is that CheXRelFormer_Local underperforms as the focus on specific anatomies limits the visual information available to the model. Given the highly fine-grained nature of this task, this result suggests that the relationship between an area and its surroundings is critical to a radiologist's perception of change, and the local Transformer cannot provide the necessary second-order information to the model. Therefore, our results show that global image-level information is crucial for accurately predicting disease change.

In addition, the absolute difference model, CheXRelFormer_AbsDiff, failed to perform the disease change classification task, indicating the importance of the proposed difference module. By incorporating the difference module and computing multi-level feature differences at multiple resolutions, CheXRelFormer learns to focus on the changes between two CXRs and to ignore irrelevant information. Our results demonstrate that multi-level feature differences are critical for improving performance in predicting disease change.

3.6 Qualitative Analysis

In Fig. 2, we visualize the model predictions from CheXRelFormer using attention rollout [1]. The produced attention maps clearly show the model's focus regions, which confirm that the model concentrated on the correct region in each image pair. CheXRelFormer can better differentiate between important and extraneous visual signals, allowing it to more accurately predict the 'no change' label. The model's ability to learn the optimal distance metric for each scale allows differentiating between relevant and irrelevant differences. In addition, analyzing multiple scales of visual features enables capturing subtle changes in pairs of CXR images, resulting in a better predictive performance for the 'improved' label. In contrast, the CheXRelFormer_AbsDiff model has difficulty in predicting both 'no change' and 'improved' labels (as shown in Fig. 3) due to the fact that images are not co-registered and exhibit several differences in their spatial or spectral characteristics – even though there was no actual change in the observed pathology.

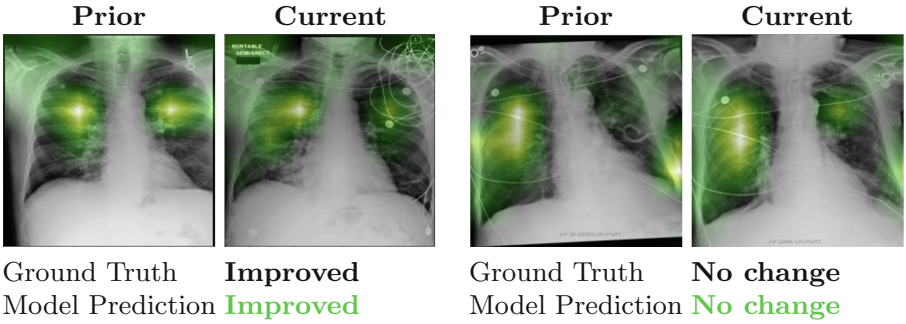

Fig. 2. Examples of model predictions obtained by `CheXRelFormer` compared against the ground-truth labels. Pathology labels LO: Lung Opacity (left) and PN: Pneumothorax (right). `CheXRelFormer` predicts the correct labels.

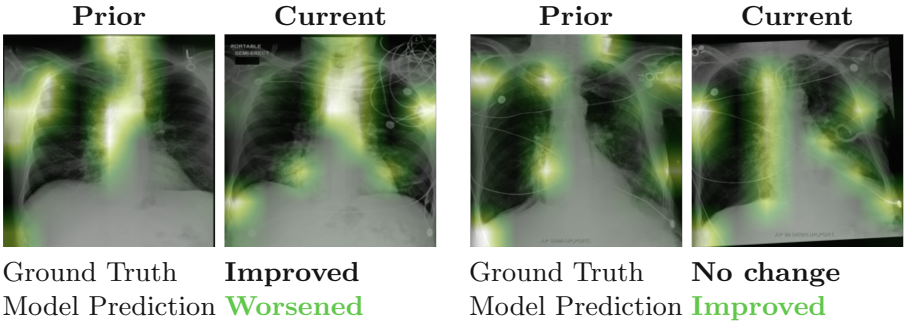

Fig. 3. Examples of model predictions obtained by `CheXRelFormer_AbsDiff` model compared against the ground-truth labels. Pathology labels LO: Lung Opacity (left) and PN: Pneumothorax (right). `CheXRelFormer_AbsDiff` struggles with differentiating between relevant and extraneous visual changes.

4 Conclusion

Monitoring disease progression is a critical aspect of patient management. This task requires skilled clinicians to carefully reason and evaluate changes in a patient's condition. In this paper, we propose `CheXRelFormer`, a hierarchical Transformer with a multi-scale difference module, trained on global image pair information to detect disease changes. Our model is inspired by the way clinicians monitor changes between CXRs, and improves the state of the art in this challenging medical imaging task. Our ablation studies show that global attention and the proposed difference module are critical components, and both help detect fine-grained changes between images. While our work shows significant progress, given the fine-grained nature of visual features that characterize findings in CXRs, disease progression remains a challenging task. In future work, we intend to include multimodal contextual information beyond the images, such

as patient history and reports, to enhance the results. `CheXRelFormer` offers a promising solution for monitoring disease progression, and future work can extend the proposed methodology to various medical imaging modalities.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190–4197 (2020)
2. Agarap, A.F.: Deep learning using rectified linear units. arXiv preprint arXiv:1803.08375 (2018)
3. Bandara, W.G.C., Patel, V.M.: A transformer-based siamese network for change detection. In: IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 207–210. IEEE (2022)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, pp. 1251–1258 (2017)
5. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
6. Huang, G., Sun, Yu., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV, pp. 646–661. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_39
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp. 448–456. PMLR (2015)
8. 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)
9. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, pp. 1–8 (2019)
10. Johnson, A.E., et al.: Mimic-iii, a freely accessible critical care database. Sci. Data **3**(1), 1–9 (2016)
11. Karwande, G., Mbakwe, A.B., Wu, J.T., Celi, L.A., Moradi, M., Lourentzou, I.: CheXRelNet: an anatomy-aware model for tracking longitudinal relationships between chest X-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I, pp. 581–591. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16431-6_55
12. Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
13. Li, M.D., et al.: Automated assessment and tracking of Covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. Radiology: Artif. Intell. **2**(4) (2020)
14. Li, M.D., et al.: Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. NPJ digital medicine **3**(1), 1–9 (2020)

15. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 269–280 (2021)
16. Liu, G., et al.: Clinically accurate chest x-ray report generation. In: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 4th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 106, pp. 249–269 (2019)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2018)
18. Park, S., et al.: Multi-task vision transformer using low-level chest x-ray feature corpus for Covid-19 diagnosis and severity quantification. Med. Image Anal. **75**, 102299 (2022)
19. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPs), pp. 8024–8035 (2019)
20. Rousan, L.A., Elobeid, E., Karrar, M., Khader, Y.: Chest x-ray findings and temporal lung changes in patients with Covid-19 pneumonia. BMC Pulm. Med. **20**(1), 1–9 (2020)
21. Tang, Y.X., et al.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. NPJ Digital Med. **3**(1), 70 (2020)
22. Wang, W., et al.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578 (2021)
23. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition, pp. 9049–9058 (2018)
24. Wu, J., et al.: Automatic bounding box annotation of chest x-ray data for localization of abnormalities. In: Proceedings of the 17th International Symposium on Biomedical Imaging (ISBI), pp. 799–803. IEEE (2020)
25. Wu, J.T., et al.: Chest imagenome dataset for clinical reasoning. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
26. Wu, J.T., Syed, A., Ahmad, H., et al.: Ai accelerated human-in-the-loop structuring of radiology reports. In: Proceedings of the Americal Medical Informatics Association (AMIA) Annual Symposium (2020)
27. Wu, J.T., et al.: Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. JAMA Netw. Open **3**(10) (2020)
28. Zhang, M., Liu, Z., Feng, J., Liu, L., Jiao, L.: Remote sensing image change detection based on deep multi-scale multi-attention siamese transformer network. Remote Sens. **15**(3), 842 (2023)