

PairAug: What Can Augmented Image-Text Pairs Do for Radiology?

Yutong Xie^{1*} Qi Chen^{1*} Sinuo Wang¹ Minh-Son To⁴ Iris Lee⁴ Ee Win Khoo⁴
 Kerolos Hendy⁴ Daniel Koh⁴ Yong Xia^{2,3} Qi Wu^{1†}

¹ Australian Institute for Machine Learning (AIML), The University of Adelaide, Australia

² School of Computer Science and Engineering, Northwestern Polytechnical University, China

³ Ningbo Institute of Northwestern Polytechnical University, China

⁴ South Australia Medical Imaging, Australia

yutong.xie678@gmail.com, {qi.chen04, qi.wu01}@adelaide.edu.au

Abstract

Current vision-language pre-training (VLP) methodologies predominantly depend on paired image-text datasets, a resource that is challenging to acquire in radiology due to privacy considerations and labelling complexities. Data augmentation provides a practical solution to overcome the issue of data scarcity, however, most augmentation methods exhibit a limited focus, prioritising either image or text augmentation exclusively. Acknowledging this limitation, our objective is to devise a framework capable of concurrently augmenting medical image and text data. We design a Pairwise Augmentation (PairAug) approach that contains an Inter-patient Augmentation (InterAug) branch and an Intra-patient Augmentation (IntraAug) branch. Specifically, the InterAug branch of our approach generates radiology images using synthesised yet plausible reports derived from a Large Language Model (LLM). The generated pairs can be considered a collection of new patient cases since they are artificially created and may not exist in the original dataset. In contrast, the IntraAug branch uses newly generated reports to manipulate images. This process allows us to create new paired data for each individual with diverse medical conditions. Our extensive experiments on various downstream tasks covering medical image classification zero-shot and fine-tuning analysis demonstrate that our PairAug, concurrently expanding both image and text data, substantially outperforms image-/text-only expansion baselines and advanced medical VLP baselines. Our code is released at <https://github.com/YtongXie/PairAug>.

1. Introduction

Vision-language pre-training (VLP) has garnered considerable attention in recent years [25], yielding substantial

benefits across a wide spectrum of downstream tasks. However, the application of VLP within the medical domain faces a complex challenge, largely attributable to the inherent requirement for extensive data. For instance, the Contrastive Language-Image Pretraining (CLIP) model [25] necessitates training on a dataset comprising 400 million image-text pairs curated from the internet. In stark contrast, the total volume of publicly accessible medical images and reports is significantly lower by several orders of magnitude. This scarcity is primarily due to privacy considerations, data acquisition challenges, and the rarity of certain diseases [8].

Data augmentation algorithms serve to address the sample size limit without altering the base model architecture, making them widely applicable across various tasks and algorithms. By expanding medical datasets using such methods, we can not only enhance the size and diversity of training datasets but also impute missing values and maintain patient privacy, thereby reducing dependence on real-world data acquisition. Many studies [7, 11, 14, 22, 24, 30] have explored image augmentation techniques designed specifically for medical image expansion, which mainly include traditional spatial transformations and morphological operations, as well as recent image synthesis techniques. These studies demonstrate significant potential in enhancing the accuracy of data-driven diagnostics and prognostics applications. Influenced by the recent advancements in Natural Language Processing (NLP), particularly the development of Large Language Models (LLMs) [1, 5, 23, 34], many studies have concentrated on the augmentation of medical textual data [6, 32, 44] such as traditional synonym replacement, random deletion and random insertion, as well as more recent methods using LLMs to generate reliable text samples.

Despite considerable advances in the field, a significant proportion of medical data augmentation algorithms remain narrowly focused, concentrating exclusively on either image or text augmentation. In the context of VLP, image and text

*Equal contribution. †Corresponding author.

data are interconnected and interdependent. Expanding data from a single modality, such as exclusively expanding images or text, fails to fundamentally enhance the information gain. This limitation arises from two main factors: Firstly, the augmented modality must retain semantic congruence with the non-augmented one, thereby constraining the scope for diversifying semantic content. For instance, an enlarged X-ray image paired with its original, unchanged report might lead to a description mismatch. Secondly, the information in the non-augmented modality remains unchanged, failing to correspond with modifications in the augmented modality. Consider a scenario where a CT scan image is altered to depict a tumour, but the associated text report remains unaltered, thus not accurately reflecting the changes in the image. Effective augmentation in VLP requires a synchronised enhancement of both image and text, ensuring coherence and maximising the information gain from paired medical image-text data. We thus argue that the simultaneous expansion of medical image and text data is paramount. By adopting this parallel augmentation strategy, we not only enlarge the dataset quantitatively but also increase data diversity at the semantic level, thereby further enhancing the model's generalisation capability and accuracy.

In this paper, we propose a Pairwise Augmentation (PairAug) approach for medical VLP. The PairAug consists of two distinct branches: the Inter-patient Augmentation (InterAug) and the Intra-patient Augmentation (IntraAug) branches, each with unique functions. These branches aim to expand paired data across inter- and intra-patient domains, maintaining a careful balance to avoid redundancy or overlap within the augmented pairs. The InterAug branch, powered by large text-to-image models, generates synthetic radiology images from plausible reports produced by the LLM. Given that these image-report pairs are entirely synthesised, they represent an ensemble of novel patient cases, most of which may not exist in the original dataset. These artificially created cases provide us with a new reservoir of data that can augment the current repository of real-world cases. Conversely, the IntraAug branch operates differently. It modifies existing images according to new reports generated by the model. This unique approach allows us to expand a multitude of new image-report pairs for each individual patient, each pair reflecting a different medical condition. The IntraAug branch thereby allows for creating a diverse dataset that better represents the variety of potential medical scenarios a patient might experience.

We further incorporate two data pruning techniques into our approach to ensure the augmented data's integrity and quality. These methods leverage a pre-trained text-image retrieval model to sift through the data, discarding noisy or irrelevant samples. This rigorous quality control process ensures that our dataset comprises the most relevant and reliable data, making it highly suitable for further analysis

and training.

We conduct medical VLP experiments using real-world pairs from the MIMIC-CXR dataset and PairAug-generated image-report pairs. The learned representations are transferred to classifying diseases under zero-shot and fine-tuning settings on these downstream tasks. Benefiting from the augmented paired data, our approach achieves mean AUCs of 88.34% and 70.79% on the ChestXpert and PadChes zero-shot protocol, respectively surpasses the advanced CheXzero by about 2.10% and 4.50% without ensemble. On RSNA fine-tuning protocol, using PairAug-generated pairs beats strong competitors like ImageNet pre-training, and advanced medical image-report pre-training competitor CheXzero. Besides, PairAug also outperforms popular image-/text-only augmentation baselines on these downstream tasks.

2. Related Works

Medical VLP Medical VLP, an extension of VLP in healthcare, aims to interpret complex medical images and associated texts. Many approaches use vision-language contrastive learning [15, 37, 40, 43, 46], leveraging naturally occurring medical image-radiology report pairs, yielding impressive results across various tasks like image classification [15, 37, 46] and image-text retrieval [15, 37, 40]. Recent advancements have seen the incorporation of Masked Auto-Encoder (MAE) model [12] from the natural images domain into medical VLP, proving beneficial [4, 47]. However, despite promising outcomes, the performance is often limited by the scarcity and diversity of real-world medical image-text pairs.

Medical Data Expansion The field of medical data expansion is an active area of research, addressing the critical issue of data scarcity in healthcare [19]. Traditional image expansion methods have primarily adopted spatial transformations such as rotation, scaling, and flipping, as well as morphological operations like cropping and padding [17, 39, 45]. Although these techniques are straightforward to implement, they may fall short of capturing the intricate variations inherent in medical images. The emergence of advanced augmentation techniques, particularly Generative Adversarial Networks (GANs), has ushered in a new era of synthetic but realistic-looking image generation. GANs and their variants have demonstrated considerable success across various medical imaging contexts [22, 24, 42], offering a richer, more diverse dataset for model training. A handful of recent works have explored the potential of Language-to-Image models in medical image expansion [3, 26, 29, 41]. They employ prompts based on medical textbooks/reports as inputs for generating images and improving the diagnostic accuracy of models trained on limited real datasets.

Traditional text expansion methods work at different granularity levels [6]: characters, words, sentences, and documents. Recent advances in LLMs, *e.g.*, PaLM [5],

LLaMA [34], and ChatGPT, have facilitated the development of more sophisticated text expansion methods. Despite substantial progress in medical data expansion, most existing strategies are confined to either image or text synthesis. This singular focus does not adequately address the persistent issue of limited real-world medical image-text pair data.

3. Proposed Method

Problem Statement To tackle the issue of limited data, we explore a novel dataset augmentation task. Considering open-vocabulary learning, we specifically aim to augment image-text pairs in an original/existing dataset Ω_o , where $(x_i, y_i) \in \Omega_o$. Here, x_i represents an image paired with corresponding text y_i (e.g., radiology reports), and n_o indicates the number of samples. The objective of dataset expansion is to generate a collection of new synthetic samples Ω_s , where $(\tilde{x}_i, \tilde{y}_i) \in \Omega_s$, to amplify the original dataset, enabling a learnable model trained on the expanded dataset $\Omega_o \cup \Omega_s$ significantly outperform a model trained solely on Ω_o . Crucially, the synthetic pairs set Ω_s should bring sufficient new and correct information to boost the model’s training.

How to Augment for Effective Expansion? To generate new image-text pairs, we leverage the capabilities of the large language model \mathcal{P} and the image synthesis model \mathcal{G} , known for their impressive text and image generation capabilities, respectively. However, the effectiveness of different sample types remains unclear. Our main insight is that the newly synthesised pairs $(\tilde{x}, \tilde{y}) \in \Omega_s$ should introduce new information compared to the original pairs $(x, y) \in \Omega_o$ while maintaining a high quality for each created pair. To achieve these, we consider two key criteria: (1) non-overlapped pairwise augmentation and (2) prioritising high-quality pairs.

Overall Pipeline As shown in Figure 1, considering the aforementioned points, we propose a framework called Pairwise Augmentation (PairAug) for expanding datasets. This framework, guided by the specified criteria, broadens the dataset through two distinct branches: Inter-patient Augmentation (InterAug) and Intra-patient Augmentation (IntraAug), thereby preventing redundancy or overlap in the augmented pairs. Moreover, we incorporate specific data pruning methods for each branch to uphold the quality of the augmented pairs. The pipeline of PairAug can be formulated as

$$\Omega_{\tilde{s}} \leftarrow \text{Pr}(\Omega_s), \text{ s.t. } \Omega_s = \{(\tilde{x}_i, \tilde{y}_i) | \tilde{x}_i = \mathcal{G}(\tilde{y}_i), \tilde{y}_i = \mathcal{P}(y_i)\}_{i=1}^{n_s}, \quad (1)$$

where $\Omega_{\tilde{s}}$ is the subset of Ω_s and n_s denotes the number of synthetic pairs in Ω_s . $\text{Pr}()$ is a pruning operation. For simplicity, we omit the input prompt for large language model \mathcal{P} . In practice, the final synthetic data set $\Omega_{\tilde{s}}$ consists of two subsets $\Omega_{\tilde{a}}$ and $\Omega_{e'}$ derived from our InterAug and IntraAug branches, respectively. In the following, we will depict how to obtain $\Omega_{\tilde{a}}$ and $\Omega_{e'}$, as well as the specific formulation of Eq. (1) in the proposed two branches.

Table 1. Reports before and after modifying based on our prompt. First row: abnormal \rightarrow normal; second row: normal \rightarrow abnormal.

Report (before)	Report (after)
<i>Mild pulmonary edema with superimposed left upper lung consolidation...</i>	<i>No pulmonary edema or lung consolidation is observed...</i>
<i>...lungs are hyperinflated though clear, cardio mediastinal silhouette is stable, ...</i>	<i>...lungs are collapsed and unclear, cardio mediastinal silhouette is unstable, ...</i>

3.1. InterAug: Inter-patient Augmentation

New Report Generated by Large Language Model (LLM) Firstly, we focus on the text domain, using ChatGPT with Azure OpenAI service to process X-ray reports. The LLM produces the resulting report based on manual instruction (prompt). To simplify the process, we aim to find a single satisfactory prompt when generating new reports, i.e., “Following is an original chest X-Ray report. Generate one possible augmentation that is limited to 50 words while conveying partial opposite meanings than the original report”.

As shown in Table 1, we take as an example that provided the input report “Mild pulmonary edema with superimposed left upper lung consolidation”, we use ChatGPT to generate an appropriately modified output report “No pulmonary edema or lung consolidation is observed” based on our written prompt. In this way, we can obtain a large number of new and diverse radiology reports. Due to the page limit, we provide more examples in the supplementary.

Inter-patient Image Generation In this part, we seek to generate images based on text without any other constraints. In this way, these pairs can be regarded as a set of new patients as they are synthetically generated and may not exist in original datasets (see the generated image in Figure 2(a)). Specifically, we base our model on Stable Diffusion, a large-scale text-to-image (T2I) latent diffusion model [28] and use the post-pretraining version on radiology datasets, namely RoentGen [3]. Formally, the process of generating new image-text pairs can be defined as

$$\Omega_a = \{(\hat{x}_i, \hat{y}_i) | \hat{x}_i = \mathcal{G}_{z \sim p_z}(\hat{y}_i, z), \hat{y}_i = \mathcal{P}_{y_i \in \Omega_o}(y_i)\}_{i=1}^{n_a}, \quad (2)$$

where \mathcal{P} refers to the LLM and \mathcal{G} is the generation model. p_z is the prior distribution for input noise variable z . n_a means the number of new image-text pairs.

Data Pruning by Semantically-aligned Informativeness To ensure the quality of the generated pairs, we introduce a data pruning method $\text{Pr}_a()$ based on a semantically-aligned score \mathcal{S}_a , which resorts to the semantic alignment abilities of CLIP [25]. Similar to [27], we rerank the samples drawn from the generation model using CLIP. Notably, as we focus on the radiology image generation in chest X-rays, we employ MedCLIP [38] pre-trained on the chest X-ray dataset

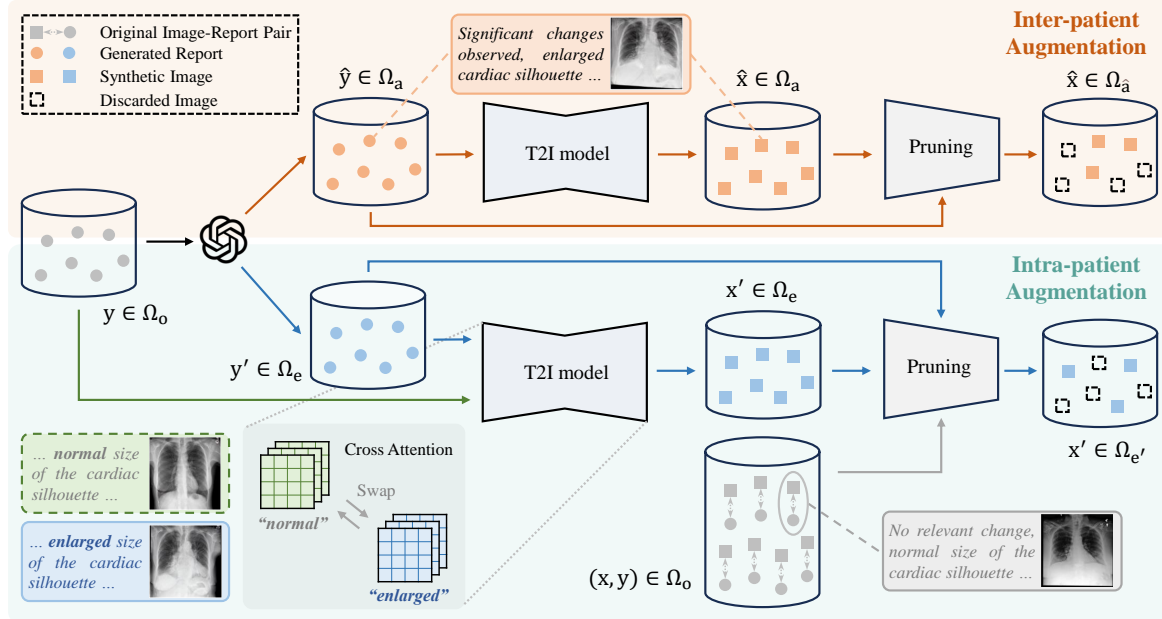


Figure 1. Overall of Pairwise Augmentation (PairAug) pipeline, consisting of two branches: Inter-patient Augmentation (InterAug) and Intra-patient Augmentation (IntraAug). In InterAug, we first generate new reports $\hat{y} \in \Omega_a$ by a large language model \mathcal{P} from original reports $y \in \Omega$. Then, we synthesise images $\hat{x} \in \Omega_a$ from the generated reports, followed by a data pruning method *w.r.t.* the semantic alignment between generated image-report pairs. As for IntraAug, we seek to generate images for the same individual but with different medical conditions. To this end, we reuse the same generation model \mathcal{G} to synthesise images but swap the cross-attention map M from the original report y with that (*i.e.*, M') from the modified report y' during the generation process. After that, we consider a data pruning method based on both synthetic pairs $(x', y') \in \Omega_e$ and original pairs $(x, y) \in \Omega$. Last, we merge $\Omega_{\hat{a}}$ and $\Omega_{e'}$ as the final synthetic paired data set $\Omega_{\hat{s}}$.

rather than the original CLIP model. Specifically, for the paired data from our InterAug, MedCLIP assigns a score based on how well the image matches the report. Then, we filter and only retain those image-report pairs that have attained scores exceeding the threshold τ . In this way, we not only ensure the semantically aligned informativeness of the extended data set but also improve the robustness of our data generation method against failures of the generation model. Mathematically, the retained data set $\Omega_{\hat{a}}$ after pruning process can be defined as

$$\Omega_{\hat{a}} \leftarrow \text{Pr}_{\tau}(\Omega_a, \tau) = \{(\hat{x}, \hat{y}) | (\hat{x}, \hat{y}) \in \Omega_a, \mathcal{S}_a(\hat{x}, \hat{y}) > \tau\}, \quad (3)$$

where $\mathcal{S}_a(\hat{x}, \hat{y})$ refers to the cosine similarity between the image feature and text feature extracted by MedCLIP’s image encoder and text encoder, respectively.

3.2. IntraAug: Intra-patient Augmentation

Besides expanding the data at the patient level through InterAug, we seek to obtain samples that capture various medical conditions for each individual. This further enhances the diversity of dataset without introducing overlap or redundancy compared to data generated by InterAug. However, due to the lack of guarantees regarding image consistency, capturing changes in each patient’s condition, even minor changes, is challenging. In Figure 2(a), the radiology images are

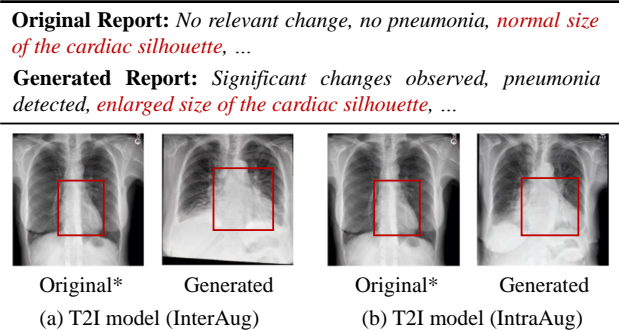


Figure 2. Images synthesised from original and generated reports by (a) the T2I model in InterAug and (b) the T2I model in IntraAug with attention map swapping, respectively. * denotes images generated from the original reports rather than the original images.

from different patients, even if both of them are semantically aligned with their reports (see the box and description in red). This contradicts our goal, as we hope to generate radiology images displaying different medical conditions for the same patient (Figure 2(b)), rather than generating conditions on another patient. To this end, we design an IntraAug to yield new intra-patient images using newly generated reports while considering the original reports as conditions. To make the whole model memory-friendly, instead of re-developing an image editing model (in which the editing quality may not

be guaranteed with limited training data), we reuse the same generation model \mathcal{G} (from the above section) to synthesise images but swap the cross-attention map from the original report. See the below section for more details.

Image Generation by Controlling Cross-Attention Maps

Inspired by the observation in [13], we seek to control the generation process by modifying the intermediate cross-attention maps M of generation model \mathcal{G} . Formally, we define the generation process as

$$\Omega_e = \{(x'_i, y'_i) | x'_i = \mathcal{E}(y_i, y'_i), y'_i = \mathcal{P}_{y_i \in \Omega_o}(y_i)\}_{i=1}^{n_e}, \quad (4)$$

where \mathcal{P} is the large language model and n_e means the total number of generated data. Here, \mathcal{E} refers to the generation process, which contains the diffusion model \mathcal{G} and an attention map swapping operation $\text{SWAP}^{(t)}(M_t, M'_t)$. This operation swaps the original cross-attention maps M_t (from the original report y) with the modified M'_t (from the modified report y') at step t of the diffusion process, *e.g.*, from “normal” to “enlarged” in Figure 1.

Mathematically, given y' , the output noisy image z'_{t-1} at the step t of diffusion processing can be calculated by

$$z'_{t-1} \leftarrow \mathcal{E}^{(t)}(y, y') = \mathcal{G}^{(t)}(y', z'_t) \{ \text{SWAP}^{(t)}(M_t, M'_t) \}. \quad (5)$$

Here, let $\mathcal{G}^{(t)}(y', z'_t) \{ \text{SWAP}^{(t)}(M_t, M'_t) \}$ represent the diffusion step where we swap the attention map M_t with the modified map M'_t , where $M_t \leftarrow \mathcal{G}^{(t)}(y, z_t)$ and $M'_t \leftarrow \mathcal{G}^{(t)}(y', z'_t)$. The noisy images z_t and z'_t are generated at the previous step from y and y' , respectively. For \mathcal{G} , a random noise $z'_T \sim p_z$ is fed at the first step T and finally yields the image $x' = z'_0$ at the last step 0. Moreover, following [13, 21], we use a softer attention map swapping method for well-controlling the degree of modification, *i.e.*,

$$\text{SWAP}^{(t)}(M_t, M'_t) := \begin{cases} M_t \leftarrow M'_t & t < \eta \\ M_t & \text{otherwise,} \end{cases} \quad (6)$$

where $\eta = 0.5$ is a timestamp hyper-parameter, specifying which step the swapping operation is used.

Data Pruning by Hybrid Consistency In this part, we assess the data from IntraAug by designing a hybrid consistency score, which consists of three consistency criteria: (i) semantic alignment \mathcal{S}_1 between input reports and corresponding images; (ii) similarity \mathcal{S}_2 between original images and generated images; (iii) the consistency of the change \mathcal{S}_3 between two images with the change between the corresponding two reports. For (i) and (ii), we directly use the MedCLIP to capture both features of reports and images, and then calculate the similarity for image-report pairs and image-image pairs, respectively. As for (iii), inspired by [10], we use the directional similarity in CLIP space. This calculates the cosine similarity between Δx and Δy , where Δx represents the differences between image features and Δy represents the differences between report features.

Due to the different magnitudes of \mathcal{S}_1 , \mathcal{S}_2 , and \mathcal{S}_3 , we seek to computer them individually and subsequently take their mean values as filtering thresholds. Formally, the data sets Ω_1 , Ω_2 and Ω_3 can be filtered by

$$\begin{cases} \Omega_1 = \{(x', y') | \mathcal{S}_1(x', y') > (\mu_1 - \epsilon)\} \\ \Omega_2 = \{(x', y') | \mathcal{S}_2(x', x) > (\mu_2 - \epsilon)\} \\ \Omega_3 = \{(x', y') | \mathcal{S}_3(\Delta x, \Delta y) > (\mu_3 - \epsilon)\}. \end{cases} \quad (7)$$

Here, Δx is the subtraction of features between x and x' while Δy is that between y and y' , where $(x', y') \in \Omega_e$ and $(x, y) \in \Omega_o$. The Ω_o is the original data set. μ_1 , μ_2 and μ_3 denote the mean value of score \mathcal{S}_1 , \mathcal{S}_2 and \mathcal{S}_3 , respectively. Besides, we introduce a hyper-parameter ϵ , making the threshold more flexible and less stringent. Finally, we obtain the augmented dataset by

$$\Omega_{e'} \leftarrow \text{Pr}_e(\Omega_e, \Omega_o, \epsilon) = \Omega_1 \cap \Omega_2 \cap \Omega_3, \quad (8)$$

where $\text{Pr}_e()$ is the pruning method with above three criteria.

3.3. Medical VLP with Generated Pairs

We amalgamate real-world and generated medical image-text pairs for our medical VLP. Our model builds upon the CheXzero framework [33], an advanced approach that proficiently exploits semantic correspondences between medical images and radiology reports for comprehensive medical data representation learning. We use the Vision Transformer [9], ViT-B/32, as the image encoder and employ a Transformer [36] with 12 layers and a width of 512 with eight attention heads for the text encoder. We initialized the self-supervised model using the pre-trained weights from OpenAI’s CLIP model [25]. After that, we apply the pre-trained weight parameters to various downstream classification tasks under zero-shot and fine-tuning settings. For the zero-shot setting, we followed the CheXzero and used a positive-negative softmax evaluation procedure on each disease for multi-label classification task. In particular, we compute logits with positive prompts (such as pneumonia) and negative prompts (that is, no pneumonia). Then, we compute the softmax between the positive and negative logits. Lastly, we keep the softmax probabilities of the positive logits as the probability of the disease in the chest X-ray. We employ the widely-used linear probing approach for the fine-tuning evaluation, in which the pre-trained image encoder is frozen, and only a randomly initialized linear classification head is trained.

4. Experiments

4.1. Implementation Details

Generation Setup Following [3], we establish a guidance scale of 4 and generate images at 512 resolution with 75 denoising steps using a PNDM noise scheduler [20]. For data pruning, we empirically set the τ and ϵ thresholds to 0.3

and 0.003, respectively, to guarantee the quality of generated image-text pairs. Finally, our PairAug generates 187,922 image-text pairs; 44,279 are from InterAug, while IntraAug produces the remaining 143,643 pairs.

Pre-training Setup We combine real-world 377k image-report pairs from the MIMIC-CXR dataset [18] and PairAug-generated image-report pairs for the pre-training. During the pre-training phase, we combine the real-world and generated pairs. We set the input image size at 224×224 and normalise each image using the training dataset’s sample mean and standard deviation. We tokenise reports using byte pair encoding with a 49,408-word vocabulary. We use the stochastic gradient descent optimiser combined with a learning rate of 0.0001, a momentum of 0.9, and a batch size of 64. The maximum training epoch is ten.

Downstream Setup We test the performance of learned VLP representations on three radiology-based downstream datasets: (1) CheXpert dataset [16] contains 191,229 frontal chest radiographs. We use its official test set for zero-shot evaluation aiming to classify each image into 5 five individual binary labels: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion; (2) PadChest dataset [2] has 193 disease image labels, including 174 radiographic findings and 19 differential diagnoses. We adopt 39,053 chest X-rays annotated by board-certified radiologists for zero-shot evaluation; (3) pneumothorax classification on the RSNA Pneumonia dataset [31] with over 29k frontal view chest radiographs, which aims to classify each radiograph into negative or positive for pneumothorax. We split the dataset into training, validation, and test sets with an 80%/10%/10% ratio. We report the area under the ROC curve (AUC), accuracy (ACC) and macro-averaged F1 score (F1) as the evaluation metrics. We put more details in the supplementary.

4.2. Comparison with State-of-the-arts

We compare our PairAug with different pre-training and data augmentation methods on three downstream datasets, including zero-shot evaluation on ChestXpert and PadChest datasets in Table 2 and linear probing evaluation on RSNA datasets under varying ratios of available labelled data (1%, 10%, and 100%) in Table 3. The compared methods include ImageNet pre-training [9], popular medical image-report pre-training approaches MGCA [15], CXR-CLIP [43] and CheXzero (our base model) [33]. Besides, we test the impact of our base model with only image augmentation (Base+AugImg), with only text augmentation (Base+AugText), with image+text augmentation (Base+AugText+AugImg), and with our paired image-text data augmentation (Base+PairAug). For AugImg, we incorporate traditional image augmentation in VLP, *i.e.*, random cropping, Gaussian Blur, and Grayscale, to expand the image data. For AugText, we employ ChatGPT to rewrite the original reports while keeping the same semantics, to expand the

text data. For AugText+AugImg, we add traditional image augmentation to the AugText setting.

Comparison with Different Augmentation Methods When comparing with image-only and text-only data augmentation methods (*i.e.*, AugImg, AugText, and their combination), Table 2 shows that our PairAug exhibits superior performance across all these scenarios.

Although AugImg and AugText contribute to performance improvement compared to the base model (CheXzero-S), their data augmentation focuses on only a single modality (either image or text), under the premise of semantic invariance, limiting their potential to enrich the information gain fundamentally. Besides, limited to the data diversity at the semantic level, it is interesting that performance is not significantly enhanced even when these two augmentation techniques are combined. In contrast, our PairAug approach generates paired medical image-text data, augmenting both image and text modalities concurrently, and importantly, without the restrictions of semantic invariance. Thus, PairAug produces a more comprehensive and contextually rich training dataset for the model, which, as demonstrated by the experimental results, improves model performance.

For ChestXpert and PadChest datasets, the performance gain achieved by PairAug over AugImg, AugText, and AugImg+AugText is significant for all metrics with a 95% confidence interval (CI). *E.g.*, for ChestXpert, PairAug achieves an average AUC of 88.34%, an average AUC of 84.97%, and an F1 score of 65.73%, which are higher than the scores achieved by AugImg (87.43, 84.07 and 62.16), AugText (87.03, 82.43 and 62.07), and AugImg+AugText (87.12, 83.69 and 62.46). The total minimum performance gain over three augmentation methods is 5.08% on the ChestXpert dataset and 3.82% on the PadChest dataset. On the RSNA dataset, PairAug consistently outperforms three augmentation methods across all proportions of available labelled data. Even with only 1% of labelled data, PairAug achieves an AUC of 85.89%, an accuracy of 83.13% and an F1 score of 73.51%. This consistent outperformance of PairAug over all these augmentation methods across different scenarios illustrates the importance and effectiveness of generating paired image-text data for medical VLP tasks.

Comparison with Different Pre-training Methods In Table 2, our PairAug, using CheXzero as a base model, sets new state-of-the-art zero-shot learning benchmarks for two key downstream datasets. Our method also beats the CheXzero model with or without using ensembles. For the fine-tuning comparisons, Table 3 shows that compared to models trained on ImageNet only, those further pre-trained on medical datasets—specifically CheXzero—consistently perform better, regardless of the labelling ratio. This demonstrates the importance of tailored pre-training for medical imaging, which has unique characteristics and requirements. Moreover, our PairAug consistently outperforms CheXzero on

Table 2. Classification results (AUC, ACC and F1) of different pre-training methods on two downstream test sets under zero-shot evaluation. For ChestXpert, the metrics all are the average of five diseases. For PadChest, the metrics all are the average of 193 diseases. Numbers within parentheses indicate 95% confidence interval (CI). ‘S’ denotes a single model. ‘E’ is the ensemble over top-ten model checkpoints.

Methods	ChestXpert			PadChest		
	AUC(%)	Acc(%)	F1(%)	AUC(%)	Acc(%)	F1(%)
MGCA (NeurIPS’22)	84.29(79.88, 88.33)	82.11(76.04, 87.00)	61.12(53.00, 69.00)	66.12(59.56, 72.05)	81.38(63.32, 91.68)	4.89(4.13, 6.60)
CXR-CLIP (MICCAI’23)	86.20(82.02, 90.08)	83.24(76.28, 88.20)	61.63(53.44, 69.73)	68.50(61.87, 74.52)	86.04(71.85, 94.13)	8.62 (6.63, 11.57)
CheXzero-E (Nat.BE’22)	88.92(84.97, 92.29)	85.75(81.84, 89.44)	66.51(58.62, 73.56)	71.24(65.52, 76.05)	84.54(70.50, 92.45)	7.36 (5.31, 9.34)
CheXzero-S (Base)	86.24(81.77, 90.16)	83.13(74.76, 88.24)	59.98(52.47, 67.18)	66.29(59.57, 72.23)	79.44(64.22, 90.69)	6.15(5.00, 8.02)
Base + AugImg	87.43(83.37, 90.96)	84.07(78.48, 88.44)	62.16(54.55, 69.29)	68.12(61.74, 73.96)	83.73(68.44, 92.56)	6.15(4.87, 8.25)
Base + AugText	87.03(82.72, 90.55)	82.43(77.24, 87.16)	62.07(54.52, 69.11)	67.35(60.82, 73.14)	81.93(66.64, 90.65)	6.41(5.13, 8.49)
Base + AugText + AugImg	87.12(82.94, 90.86)	83.69(77.96, 88.40)	62.46(54.64, 69.99)	69.23(63.01, 74.48)	83.39(68.37, 92.46)	5.91(4.63, 8.10)
Base + PairAug (Ours)	88.34(84.31, 91.84)	84.97(80.00, 88.76)	65.73(57.65, 73.11)	70.79(64.90, 75.97)	84.90(71.63, 93.18)	7.51(5.99, 9.95)
Base + PairAug-E (Ours)	89.97 (86.00, 93.27)	86.21 (81.60, 89.92)	67.78 (59.72, 75.07)	72.51 (66.36, 77.77)	86.51 (72.59, 93.77)	7.67 (6.28, 10.08)

Table 3. Classification results of different pre-training methods on RSNA test sets under different ratios of available labelled data.

Methods	1%			100%		
	AUC	Acc	F1	AUC	Acc	F1
Rand	61.48	77.62	43.70	77.67	79.01	59.89
ImageNet	74.98	77.51	48.53	83.63	81.60	69.19
CheXzero-S (Base)	83.45	80.5	70.05	86.99	83.56	75.08
Base + AugImg	84.80	82.02	72.28	87.84	84.67	76.41
Base + AugText	84.42	81.76	71.83	87.40	83.82	75.31
Base + AugText + AugImg	84.79	81.93	71.79	87.78	84.68	76.22
Base + PairAug (Ours)	85.89	83.13	73.51	88.63	85.23	77.10

the RSNA dataset whenever using different proportions of labelled medical data for fine-tuning. These superior results further suggest the effectiveness of our data augmentation strategy. By augmenting image and text data concurrently, PairAug enhances the generalisation performance of these downstream tasks. This is particularly beneficial when only a limited amount of labelled data is available.

4.3. Ablation Study

To test the performance of each component in PairAug, we conduct an ablation study for InterAug, IntraAug and two data pruning strategies in Table 4. We gradually add each component to the base model to observe downstream zero-shot performance trends on ChestXpert and PadChest datasets. First, incorporating the InterAug/IntraAug module leads to a limited, even decreased, performance gain on both datasets. This can be attributed to the generated data that possibly introduce noise or inconsistencies, causing the model to underperform. The subsequent addition of both data pruning mechanisms results in a noticeable improvement in average performance (+3.05%) compared with the base model, demonstrating the value of this component in filtering out low-quality synthesised data and reducing noise in the training set. Finally, when we incorporate all components into the framework, the performance on both datasets reaches its peak (+5.06%). It suggests that jointly two branches can provide more diverse data for pre-training, further enhancing the model’s generalisation ability.

Table 4. Ablation study. Δ is average performance gain compared to Base. ‘ChestX.’: CheXpert dataset. ‘PadC.’: PadChest dataset.

Base	Ablations				ChestX. AUC	PadC. AUC	Δ
	InterAug	$P r_a$	IntraAug	$P r_e$			
✓					86.24	66.29	-
✓	✓				84.91	66.03	-1.59
✓	✓	✓			87.36	68.22	3.05
✓			✓		85.79	67.88	1.14
✓			✓	✓	87.30	68.28	3.05
✓	✓	✓	✓	✓	88.34	69.21	5.06

4.4. Information Gain from Synthetic Data

To illustrate the spread of the synthesised data, we randomly sample 5,000 image-report pairs from the augmented datasets created using the IntraAug and InterAug methods and the original MIMIC CXR dataset. We then extract the embeddings for both images and texts from these pairs. To visualise how the synthesised data compares to the original ones, we used t-SNE [35] to map the high-dimensional embeddings to a two-dimensional plane, as shown in Figure 3. The t-SNE visualisation of synthesised and real data distributions suggests that our PairAug produce new, realistic variations that complement the existing MIMIC CXR dataset. The distributions are close in the embedding space yet with limited overlap, indicating an expansion of the dataset’s diversity without diverging from authentic medical cases. This additional variety in the synthesised data could enhance model generalisation, as it introduces unique, realistic scenarios for robust deep learning training.

4.5. Qualitative Results

Figures 4 and 5 provide a visualisation of the results of our approach, revealing three interesting observations. First, ChatGPT can effectively facilitate text augmentation. Given different prompts, ChatGPT can generate diverse reports and mimic the generation of novel clinical cases, thus gaining new information. However, note that ChatGPT occasionally outputs informal terms, such as “homogeneous clearance”. This highlights a potential area for future improvements, such as developing more refined prompts or filtering mecha-

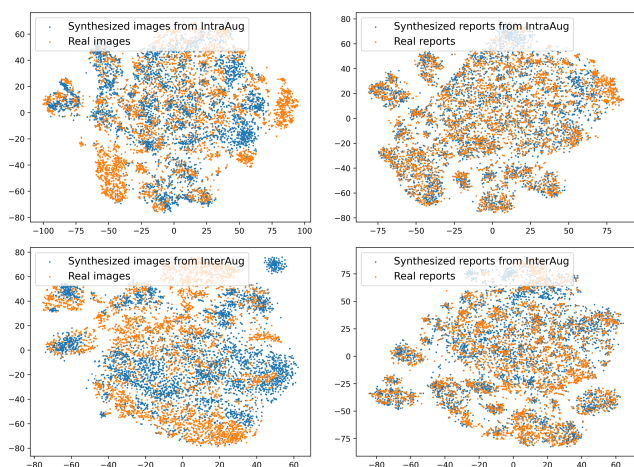


Figure 3. T-SNE visualisation of image/report embeddings, comparing synthesised data from IntraAug and InterAug methods against real data from the MIMIC CXR dataset.

nisms to remove these casual terms. Secondly, our PairAug shows its capacity to generate medical images with a strong semantic alignment with the corresponding reports, as denoted by the red boxes in images and red words in reports. Third, Figure 5 provide more evidence that our IntraAug can generate radiology images representing varying medical conditions for a specific patient. For instance, the manifestation of pulmonary edema in both lungs could potentially indicate a worsening condition compared to the initial stage, where the pulmonary edema is only present in the left lung. Some failure cases are shown in the supplementary.

4.6. Human Radiologists Study

We conducted a Visual Turing Test with medical experts, consisting of three distinct tasks to validate the authenticity of the image-text pairs generated by the PairAug method. Two radiology residents participated in this test. Task 1 involved 50 patient data sets, including real images and PairAug synthetic images. Task 2 involved 50 patient data sets, including real reports and PairAug synthetic reports. For these two Tasks, the experts were asked to identify each image/report as “real” or “synthetic/unsure”. Task 3 involved 50 patient data sets, including synthetic pairs produced by PairAug. For this task, the experts were asked to assess the semantic alignment between the image and its corresponding report, categorising each pair as “good”, “poor”, or “unsure”.

The results reflect the realism of medical image-report pairs generated by PairAug. With an average accuracy of 61% and 50% in Tasks 1 and 2, respectively, experts find distinguishing between real and synthetic images/reports challenging, suggesting a high degree of fidelity in the synthetic data. In addition, the average performance of 61% in Task 3 suggests that although the coherence between modalities is generally convincing, there is room for further improvement.

Limitation and Future Work The performance of PairAug

Real-world reports

Moderate unchanged cardiomegaly, no edema, the lungs are well expanded and clear, the mediastinal silhouette and hilar contours are normal, no pleural effusion or pneumothorax is present.

Edited reports by ChatGPT Generated images by InterGen

Severe cardiomegaly with significant changes noted, edema present, lungs are poorly expanded and hazy, mediastinal silhouette and hilar contours are normal, no pleural effusion and pneumothorax.

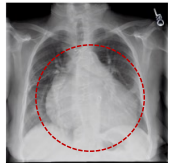



Figure 4. Radiology report before and after editing by ChatGPT and the corresponding images generated by our InterAug. We highlight the specific areas in the radiology image with red bounding boxes and the corresponding descriptions in reports with the same colour.

Real-world reports X-ray images

..., there is mild asymmetric pulmonary edema in the left lung single upright view of the chest provided there is no focal consolidation effusion or pneumothorax, there is mild pulmonary vascular congestion and mild asymmetric pulmonary edema in the left lung, ...



Edited reports by ChatGPT Generated images by IntraEdit

..., there is significant symmetric pulmonary edema in both lungs, multiple views of the chest provided, there is focal consolidation effusion and pneumothorax, there is severe pulmonary vascular congestion and symmetric pulmonary edema, ...




Figure 5. Radiology image-report pair synthesised via IntraAug.

relies heavily on the quality of the generation model employed. Here, we use the RoentGen model, pre-trained only on chest X-ray datasets, where the underlying training data is relatively biased and lacks diversity. In future work, our goal is to train a model capable of generating radiology images from reports across various body parts, expanding our approach to a wider range of medical scenarios.

5. Conclusion

In this paper, we propose an approach called PairAug to address the challenge of acquiring paired image-text datasets in radiology. PairAug consists of two branches: InterAug and IntraAug. InterAug generates synthetic radiology images paired with plausible reports, creating new patient cases, while IntraAug focuses on generating diverse paired data for each individual. We employ data pruning techniques to ensure high-quality data. Experimental results across various tasks show that PairAug outperforms baseline methods that only focus on either image or text expansion.

Acknowledgements Yutong Xie was supported by the Centre for Augmented Reasoning (CAR) project. Yong Xia was supported in part by the National Natural Science Foundation of China under Grants 62171377, and in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06).

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.*, 33:1877–1901, 2020. [1](#)
- [2] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. [6](#)
- [3] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Polacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. [2](#), [3](#), [5](#)
- [4] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *MICCAI*, pages 679–689, 2022. [2](#)
- [5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. [1](#), [2](#)
- [6] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, 2023. [1](#), [2](#)
- [7] Onat Dalmaç, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. [1](#)
- [8] Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *ICML*, pages 5378–5396, 2022. [1](#)
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021. [5](#), [6](#)
- [10] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, pages 1–13, 2022. [5](#)
- [11] John T Guibas, Tejal S Virdi, and Peter S Li. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*, 2017. [1](#)
- [12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16000–16009, 2022. [2](#)
- [13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. [5](#)
- [14] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. [1](#)
- [15] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Int. Conf. Comput. Vis.*, pages 3942–3951, 2021. [2](#), [6](#)
- [16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Siyvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. [6](#)
- [17] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, pages 203–211, 2021. [2](#)
- [18] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, pages 1–8, 2019. [6](#)
- [19] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023. [2](#)
- [20] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. [5](#)
- [21] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. *arXiv preprint arXiv:2303.04761*, 2023. [5](#)
- [22] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *MICCAI*, pages 417–425. Springer, 2017. [1](#), [2](#)
- [23] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. [1](#)
- [24] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 6839–6853, 2021. [1](#), [2](#)
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. [1](#), [3](#), [5](#)
- [26] Sajith Rajapaksa, Jean Marie Uwabeza Vianney, Renell Castro, Farzad Khalvati, and Shubhra Aich. Using large text-to-image models with structured prompts for skin disease identification: A case study. *arXiv preprint arXiv:2301.07178*, 2023. [2](#)
- [27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. [3](#)

- [28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3
- [29] Hojjat Salehinejad, Shahrokh Valaei, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 990–994. IEEE, 2018. 2
- [30] Shitong Shao, Xiaohan Yuan, Zhen Huang, Ziming Qiu, Shuai Wang, and Kevin Zhou. Diffuseexpand: Expanding dataset for 2d medical image segmentation using diffusion models. *arXiv preprint arXiv:2304.13416*, 2023. 1
- [31] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019. 6
- [32] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*, 2023. 1
- [33] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. 5, 6
- [34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 3
- [35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [37] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Adv. Neural Inform. Process. Syst.*, 2022. 2
- [38] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 3
- [39] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *MICCAI*, pages 171–180. Springer, 2021. 2
- [40] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI*, pages 2982–2990, 2022. 2
- [41] Xingyi Yang, Nandiraju Gireesh, Eric Xing, and Pengtao Xie. Xraygan: Consistency-preserving generation of x-ray images from radiology reports. *arXiv preprint arXiv:2006.10552*, 2020. 2
- [42] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. 2
- [43] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 2, 6
- [44] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. *arXiv preprint arXiv:2303.16756*, 2023. 1
- [45] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8543–8553, 2019. 2
- [46] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, pages 32–40, 2022. 2
- [47] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *Int. Conf. Learn. Represent.*, 2023. 2