Diversity-Preserving Chest Radiographs Generation from Reports in One Stage

Zeyi Hou[1], Ruixin Yan[2], Qizheng Wang[2], Ning Lang[2],
and Xiuzhuang Zhou[1](✉)

[1] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
xiuzhuang.zhou@bupt.edu.cn

[2] Department of Radiology, Peking University Third Hospital, Beijing, China

Abstract. Automating the analysis of chest radiographs based on deep learning algorithms has the potential to improve various steps of the radiology workflow. Such algorithms require large, labeled and domain-specific datasets, which are difficult to obtain due to privacy concerns and laborious annotations. Recent advances in generating X-rays from radiology reports provide a possible remedy for this problem. However, due to the complexity of medical images, existing methods synthesize low-fidelity X-rays and cannot guarantee image diversity. In this paper, we propose a diversity-preserving report-to-X-ray generation method with one-stage architecture, named DivXGAN. Specifically, we design a domain-specific hierarchical text encoder to extract medical concepts inherent in reports. This information is incorporated into a one-stage generator, along with the latent vectors, to generate diverse yet relevant X-ray images. Extensive experiments on two widely used datasets, namely Open-i and MIMIC-CXR, demonstrate the high fidelity and diversity of our synthesized chest radiographs. Furthermore, we demonstrate the efficacy of the generated X-rays in facilitating supervised downstream applications via a multi-label classification task.

Keywords: Chest X-ray generation · Radiology report · Generative adversarial networks · One-stage architecture

1 Introduction

Chest radiography is currently the most common imaging examination, playing a crucial role in epidemiological studies [3] and clinical diagnosis [10]. Nowadays, the automated analysis of chest X-rays using deep learning algorithms has attracted increasing attention due to its capability of significantly reducing the workload of radiologists and expediting clinical practice. However, training deep learning models to achieve expert-level performance on various medical imaging tasks requires large, labeled datasets, which are typically difficult to obtain due to data privacy and workload concerns. Developing generative models for high-fidelity X-rays that faithfully represent various medical concepts in radiology reports presents a possible remedy for the lack of datasets in the medical

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 482–492, 2023.
https://doi.org/10.1007/978-3-031-43904-9_47

domain [5,15]. This approach may substantially improve traditional supervised downstream tasks such as disease diagnosis [10] and medical image retrieval [8].

Generating chest radiographs based on radiology reports can be thought of as transforming textual input into visual output, while current methods typically rely on text-to-image generation in computer vision. The fidelity and diversity of synthesized images are two major qualities of generative models [1], where fidelity means the generated images should be close to the underlying real data distribution, while diversity means the output images should ideally cover a large variability of real-world images. Recently, several works have been extensively proposed to generate high-fidelity images using GANs according to text prompts. StackGAN [25] stacked multiple generators and discriminators to gradually increase the resolution of the generated images. AttnGAN [23] synthesized images with fine-grained details by introducing a cross-modal attention mechanism between subregions of images and relevant words. DM-GAN [27] introduced a memory module to refine fuzzy image contents caused by inferior initial images in stacked architecture. MirrorGAN [16] reconstructed the text from the synthesized images to preserve cross-domain semantic consistency. Despite the progress of text-to-image generation in the general domain, generating X-rays from radiology reports remains challenging in terms of word embedding, handling the linguistic structure in reports, cross-modal feature fusion, etc.

The first work to explore generating chest X-rays conditioned on clinical text prompts is XrayGAN [24], which synthesized high-fidelity images in a progressive way. Additionally, XrayGAN [24] proposed a hierarchical attentional text encoder to handle the linguistic structure of radiology reports, as well as a pre-trained view-consistency network to constrain the generators. Although impressive results have been presented, three problems still exist: 1) The progressive generation stacks multiple generators of different scales trained separately in an adversarial manner (see Fig. 1), where visual features of different scales are difficult to be fused uniformly and smoothly, making the final refined X-rays look like a simple combination of blurred contours and some details (see Fig. 2). 2) A high proportion of reconstruction loss and a pre-trained view-consistency network are used at each layer for the convergence of training stacked generators, which severely limits the diversity of generated X-rays (only one chest radiograph can be generated from one report). 3) The word vectors in [24] are based on the order in which words appear in the vocabulary, ignoring the information presented in medical-specific structures and the internal structure of words. More recently, another line of works [1,2] investigated adapting pre-trained visual-language foundational models to generate chest X-rays. However, transferring the diffusion models [18] trained with large multi-modal datasets to the medical imaging domain typically has high computational requirements.

In this paper, we propose a new report-to-X-ray generation method called DivXGAN to address the above issues. As illustrated in Fig. 1, DivXGAN allows for the synthesis of various X-rays containing relevant clinical concepts from a single report. The following contributions are made: 1) Inspired by the one-stage architecture [21], we propose to directly synthesize high-fidelity X-rays without

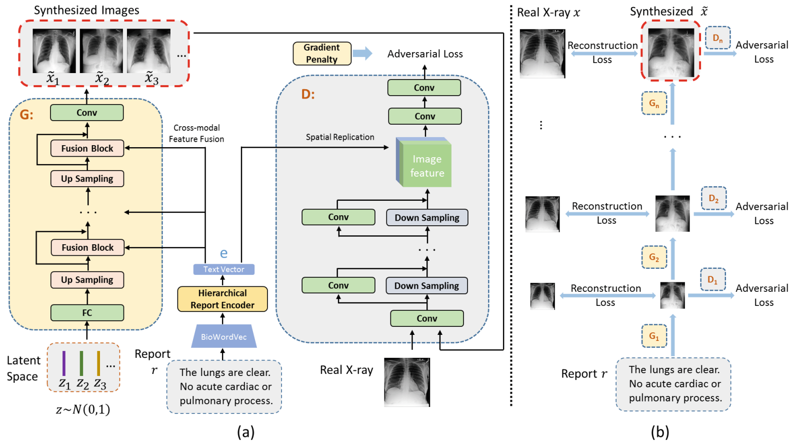

Fig. 1. (a) Overview of the proposed diversity-preserving report-to-X-ray generation with one-stage architecture. (b) Existing report-to-X-ray generative model [24]. DivXGAN discards the stacked structure with strong constraints and incorporates necessary variability via latent vectors, thus synthesizing X-rays with high fidelity and diversity.

entangling different generators. 2) We discard the pixel-level reconstruction losses and introduce noise vectors in the latent space of the generator to provide the variability, thus allowing the diversity of generated chest radiographs. 3) Lastly, we design a domain-specific hierarchical text encoder to represent semantic information in reports and perform multiple cross-modal feature fusions during X-ray generation. We demonstrate the superiority of our method on two benchmark datasets and a downstream task of multi-label classification.

2 Method

Let $\mathcal{X}$ and $\mathcal{Z}$ denote the image space and the low-dimensional latent space, respectively. Given a training set $\{x_i, r_i\}_{i=1}^{N}$ of N X-ray images, each of which x_i is associated with a radiology report r_i. The task of report-to-X-ray generation aims to synthesize multiple high-fidelity chest radiographs $\left\{\tilde{x}_i^{(j)} \in \mathcal{X}\right\}_{j=1,2,\ldots,}$ from the corresponding report r_i and latent noises $\{z_j \in \mathcal{Z}\}_{j=1,2,\ldots,}$. The generative models are expected to produce X-rays with high fidelity and diversity, so as to be used for data augmentation of downstream applications.

2.1 Fidelity of Generated X-Rays

One-Stage Generation. Existing generative method [24] uses a stacked structure to progressively synthesize high-fidelity X-rays. The stacked structure stabilizes the training of GANs but induces entanglements between generators trained separately in an adversarial way at different scales, resulting in fuzzy or discontinuous images. We draw inspiration from the one-stage architecture [21] and propose to directly generate high-fidelity X-rays using a single pair of generator G and discriminator D. The network architecture of our method is illustrated in Fig. 1. The generator contains many up-sampling layers to increase the resolution of the synthesized X-ray $\tilde{x}_i$, while the corresponding discriminator also requires many down-sampling operations to compute the adversarial loss. To stabilize the training of deep networks in this design, we introduce residual connections [6] in both the generator and the discriminator.

Distill and Incorporate Report Knowledge. Semantic information and medical concepts in radiology reports should be fully interpreted and incorporated into visual features to reduce the distance between the generated data distribution and the real data distribution, thereby improving fidelity. We design a medical domain-specific text encoder with hierarchical structure to extract the embeddings of the free-text reports. At the word level, each sentence is represented as a sequence of T word tokens, plus a special token $[SENT]$. We embed each word token w_t with an embedding matric W_e, i.e., $e_t = W_e w_t$. Unlike previous work [24] that use one-hot embedding, we initialize our word embeddings with the pre-trained biomedical embeddings BioWordVec [26], which can capture the semantic information in the medical domain and the internal structure of words. Then, we use a Transformer encoder with positional embedding to capture the contextual information of each word and aggregate the holistic representations of the sentence into the embedding of the special token $e_{[SENT]}$:

$$e_{[SENT]} = TrsEncoder\left(\left\{e_{[SENT]}, e_1, e_2, ..., e_T\right\}\right) \tag{1}$$

At the sentence level, a report consists of a sequence of S sentences, each of which is represented as $e^{(i)}_{[SENT]}$ using the word-level encoder described above. We also utilize a Transformer to learn the contextual importance of each sentence and encode them into a special token embedding $e_{[REPO]}$, which serves as the holistic representation of the report:

$$e_{[REPO]} = TrsEncoder\left(\left\{e_{[REPO]}, e^{(1)}_{[SENT]}, e^{(2)}_{[SENT]}, ..., e^{(S)}_{[SENT]}\right\}\right) \tag{2}$$

Moreover, we perform cross-modal feature fusion after each up-sampling module of the generator (see Fig. 1), to make the synthesized X-rays more faithful to the report. The fusion block contains two channel-wise Affine transformations and two ReLU layers. The scaling and shifting parameters of each Affine transformation are predicted by two MLPs (Multilayer Perceptron), using the vector $e_{[REPO]}$ as input. The Affine transformation expands the representation space of the generator G, allowing for better fusion of features from different modalities.

2.2 Diversity of Generated X-Rays

A radiology report is a medical interpretation of the corresponding chest radiograph, describing the clinical information included and assessing the patient's physical condition. Reports that describe chest radiographs of different patients with similar physical conditions are often consistent. Ideally, multiple X-ray images with the same health conditions could be generated from a single report, only with some differences in irrelevant factors such as body size, etc.

To this end, we omit the pixel-wise reconstruction loss and introduce noise vectors z in the latent space $\mathcal{Z}$ as one of the inputs to our one-stage generator, thereby providing the model with the necessary variability to ensure the diversity of synthesized X-rays. In this case, the generator G maps the low-dimensional latent space $\mathcal{Z}$ into a specific X-ray image space $\mathcal{X}_r$, conditioned on the report vector $e^i_{[REPO]}$:

$$\tilde{x}_i^{(j)} \leftarrow G\left(z_j, e^i_{[REPO]}\right),\ j = 1, 2, ... \tag{3}$$

where $\tilde{x}_i^{(j)}$ denotes the j-th synthesized X-ray from the i-th report r_i. The noise vector $z_j \in \mathcal{Z}$ follows a standard multivariate normal distribution $\mathcal{N}(0, I)$. In this way, given a radiology report, noise vectors can be sampled to generate various chest X-rays matching the medical description in the report.

2.3 Learning Objectives and Training Process

Since DivXGAN uses a one-stage generator to directly generate high-fidelity chest radiographs, only one level of generator and discriminator needs to be alternately trained. The discriminator D outputs a scalar representing the probability that the input X-ray came from the real dataset and is faithful to the input report. There are three kinds of inputs that the discriminator can observe: real X-ray with matching report, synthesized X-ray with matching report, and real X-ray with mismatched report. The discriminator $D\left(x, e_{[REPO]}; \theta_d\right)$ is trained to maximize the probability of assigning the report vector $e_{[REPO]}$ to the corresponding real X-ray x_i, while minimizing the probability of the other two kinds of inputs. Due to multiple down-sampling blocks and residual connections, we employ the hinge loss [13] to stabilize the training of D:

$$\begin{aligned} L_D = & \mathbb{E}_{x \sim p_{data}}\left[max\left(0, 1 - D\left(x, e_{[REPO]}\right)\right)\right] \\ & + \mathbb{E}_{G\left(z, e_{[REPO]}\right) \sim p_g}\left[max\left(0, 1 + D\left(G\left(z, e_{[REPO]}\right), e_{[REPO]}\right)\right)\right] \\ & + \mathbb{E}_{x \sim p_{mis}}\left[max\left(0, 1 + D\left(x, e_{[REPO]}\right)\right)\right] \end{aligned} \tag{4}$$

where p_{data}, p_g and p_{mis} denote the data distribution, implicit generative distribution (represented by G) and mismatched data distribution, respectively.

The generator $G\left(z, e_{[REPO]}; \theta_g\right)$ builds a mapping from the latent noise distribution to the X-ray image distribution based on the correlated reports, fooling the discriminator to obtain high scores:

$$L_G = -\mathbb{E}_{G\left(z, e_{[REPO]}\right) \sim p_g}\left[D\left(G\left(z, e_{[REPO]}\right), e_{[REPO]}\right)\right] \tag{5}$$

It is worth noting that the parameters θ_t of the text encoder in Eqs. (1) and (2) are learned simultaneously during the training of the generator G.

3 Experiments and Results

3.1 Datasets and Experimental Settings

We use two public datasets, namely Open-i [4] and MIMIC-CXR [9], to evaluate our generative model. The public subset of Open-i [4] consists of 7,470 chest X-rays with 3,955 associated reports. Following previous works [14,24], we select studies with two-view X-rays and a report, then end up with 2,585 such studies. As for MIMIC-CXR [9], which contains 377,110 chest X-ray images and 227,827 radiology reports, for a fair comparison with alternative methods, we also conduct experiments on the p10 subset with 6,654 cases to verify the effectiveness of our approach. Moreover, we adopt the same data split protocol as used in XRayGAN [24] for these two datasets, where the ratio of the train, validation, and test sets are 70%, 10%, and 20%. For consistency, we follow the set-up of XRayGAN [24] to focus on two major sections in each free-text radiology report, namely the "findings" section and the "impression" section.

Our network is trained from scratch using the Adam [11] optimizer with β_1=0.0 and β_2=0.9. The learning rates for G and D are set to 0.0001 and 0.0004, respectively, according to the Two Timescale Update Rule [7]. The hidden dimension of the Transformer in the text encoder is 512. The noise vector z in the latent space is sampled from a standard multivariate normal distribution with a dimension of 100. The resolution of synthesized X-rays is 512 × 512. We implemented our method using PyTorch 1.7 and two GeForce RTX 3090 GPUs. We use Inception Score (IS) [19] and Fréchet Inception Distance (FID) [7] to assess the fidelity and diversity of the synthesized X-rays. Typically, IS and FID are calculated using an Inception-V3 model [20] pre-trained on ImageNet, which might fail in capturing relevant features of the chest X-ray modality [12]. Therefore, we calculate these metrics from the intermediate layer of a pre-trained CheXNet [17]. Higher IS and lower FID indicate that the generated X-rays are more similar to the original X-rays. In addition, we also calculate the pairwise Structural Similarity Index Metric (SSIM) [22] to evaluate the diversity of X-rays generated by different methods. A lower SSIM indicates a smaller structural similarity between images, which combined with a low FID, can be interpreted as higher generative diversity [1].

3.2 Results and Analysis

We compare our approach with several state-of-the-art methods based on generative adversarial networks, including text-to-image generation: StackGAN [25], AttnGAN [23], and report-to-X-ray generation: XRayGAN [24]. The performance of different approaches on the test sets of Open-i and MIMIC-CXR is shown in Table 1. We can observe that XRayGAN [24] achieves better IS and FID than other text-to-image generation baselines. This is because the pixel-wise reconstruction loss imposes strong constraints on the stacked generators at different scales to help avoid generating insensible images. However, this strong constraint severely reduces the diversity of generated X-rays, resulting in the

Table 1. Performance of different methods on Open-i and MIMIC-CXR.

Method	Open-i			MIMIC-CXR		
	IS↑	FID↓	SSIM↓	IS↑	FID↓	SSIM↓
StackGAN [25]	1.043	243.4	0.138	1.063	245.5	0.212
AttnGAN [23]	1.055	226.6	0.171	1.067	232.7	0.231
XRayGAN [24]	1.081	141.5	0.343	1.112	86.15	0.379
DivXGAN	**1.105**	**93.42**	**0.114**	**1.119**	**62.06**	**0.143**

worst SSIM. Our method consistently outperforms other alternatives by achieving both the lowest FID and the lowest SSIM, which means the generated X-rays have better fidelity and diversity. The reason lies in that the latent noise vectors impose the necessary variation factor, and the one-stage generation process and multiple cross-modal feature fusions improve the image quality.

We conduct ablation experiments to quantify the impact of different components in DivXGAN. The results in Table 2 show that our one-stage architecture definitely improves performance due to better cross-modal feature fusion. The domain-specific encoder outperforms the hierarchical attentional encoder [24], regardless of the backbone structure, indicating the advantage of the domain-specific embedding matrix, especially for medical reports with very rare and specific vocabulary. The comparison of SSIM for different components demonstrates that the latent vector input preserves the diversity of synthesized X-rays.

Visualization of chest X-rays synthesized from a report using different methods is shown in Fig. 2. As we can see, the X-rays generated by text-to-image baselines are very coarse, and even the outline of the hearts can be barely recognized. XRayGAN [24] alleviates the blur and generates X-rays with relatively obvious chest contours, because of the strong constraints that prevent the generative model from producing abnormal samples. However, this strongly constrained approach still fails to preserve a clear outline of the heart and ribs, due to the entanglements between generators introduced by the stacked architecture. In particular, the lack of variability in the strong constraints results in only one X-ray being generated per report. Our method prevails over other alternatives as the generated X-rays are obviously clearer and more realistic, and even generate annotation information in the top right corner (seen in almost all samples). This phenomenon indicates the efficacy of the one-stage generation and multiple cross-modal feature fusion in generating high-fidelity X-rays. Furthermore, our method can generate various X-rays from one report, each of which manifests the relevant clinical findings. For example, the regions marked by red arrows in Fig. 2 show that our method can synthesize various different X-rays matching the clinical finding "Cardiomegaly" described in the report. Although our model is capable of generating X-rays based on radiology reports, due to the complexity of medical images, the synthesized X-rays have a limited range of gray-scale values and may not effectively capture high-frequency information such as subtle lung markings.

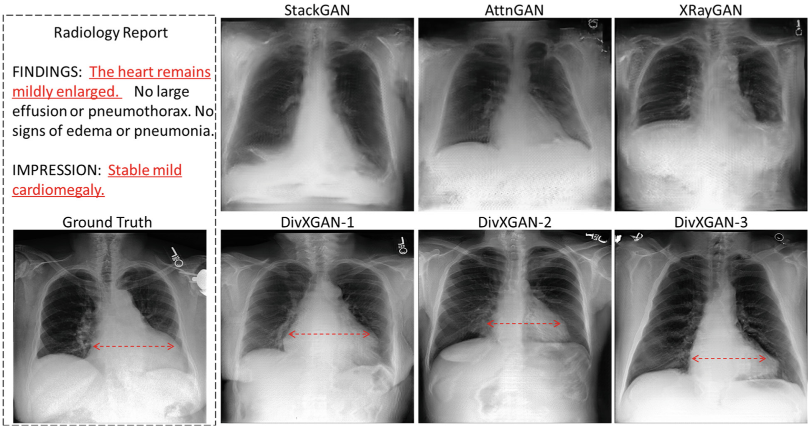
StackGAN
AttnGAN
XRayGAN
Radiology Report
FINDINGS: The heart remains mildly enlarged. No large effusion or pneumothorax. No signs of edema or pneumonia.
IMPRESSION: Stable mild cardiomegaly.
Ground Truth
DivXGAN-1
DivXGAN-2
DivXGAN-3

Fig. 2. X-ray images generated by different methods from a radiology report, where the red arrows mark the clinical finding "Cardiomegaly" described by the underlined sentences in the report. (Color figure online)

Table 2. Ablation study on Open-i (One-stage structure has latent vector input).

Methods	FID ↓	SSIM ↓
Stack w/ Hia-encoder	141.5	0.343
Stack w/ Med-encoder	139.3	0.358
One-stage w/ Hia-encoder	98.91	0.151
One-stage w/ Med-encoder	**93.42**	**0.114**

Table 3. Classification performance of a DenseNet-121 trained with various splits.

Experiment	Training Data		AUROC
	Real	Synthetic	
Real	5k	–	0.683
Synth	–	5k	0.664(↓0.019)
Real+Synth	5k	5k	0.714(↑0.031)

Furthermore, the ethical implications associated with the misuse of generated X-rays are significant. They should not be solely relied upon for clinical decision-making or used to train inexperienced medical students as radiologists. While the direct use of generated X-rays in clinical studies or medical training programs may not be appropriate, they can still serve valuable purposes in research, data augmentation, and other potential applications within the medical field. Here, we train a DenseNet-121 from scratch using various splits of real data (from MIMIC-CXR) and synthesized data (from DivXGAN) to demonstrate that the generated chest radiographs can be used for data augmentation when training downstream tasks. The task is a multi-label classification of four findings ("Cardiomegaly", "Consolidation", "Pleural Effusion" and "No Findings"). We randomly sample 5k real images with corresponding reports from the test set. These 5k real reports are input into our generative model, generating one image per report using a single latent vector, resulting in 5k generated images. When training the multi-label classifier, both real and generated images undergo the same general data augmentation techniques, such as rotation and scaling. As shown in Table 3, compared to the baseline trained exclusively on 5k real X-rays, the AUROC of the classifier trained exclusively on 5k synthesized X-rays drops by 0.019. However, training the classifier with 5k real X-rays and 5k generated X-rays improves AUROC by 0.031, suggesting that the synthesized X-rays can augment real data for supervised downstream applications.

4 Conclusion

In this paper, we have devised a diversity-preserving method for high-fidelity chest radiographs generation from the radiology report. Different from state-of-the-art alternatives, we propose to directly synthesize high-fidelity X-rays using a single pair of generator and discriminator. A domain-specific text encoder and latent noise vectors are introduced to distill medical concepts and incorporate necessary variability into the generation process, thus generating X-rays with high fidelity and diversity. We show the capability of our generative model in data augmentation for supervised downstream applications. Investigation of capturing high-frequency information of X-rays in generative models can be an interesting and challenging direction of future work.

Acknowledgements. This work is partially supported by the National Natural Science Foundation of China under grants 61972046, the Beijing Natural Science Foundation under grant Z190020, and the Proof of Concept Program of Zhongguancun Science City and Peking University Third Hospital under grant HDCXZHKC2022202.

References

1. Chambon, P., et al.: RoentGen: vision-language foundation model for chest X-ray generation. arXiv preprint arXiv:2211.12737 (2022)
2. Chambon, P., Bluethgen, C., Langlotz, C.P., Chaudhari, A.: Adapting pretrained vision-language foundational models to medical imaging domains. arXiv preprint arXiv:2210.04133 (2022)
3. Cherian, T., et al.: Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies. Bull. World Health Organ. **83**, 353–359 (2005)
4. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
5. Ganesan, P., Rajaraman, S., Long, R., Ghoraani, B., Antani, S.: Assessment of data augmentation strategies toward performance improvement of abnormality classification in chest radiographs. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 841–844. IEEE (2019)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

8. Huang, P., Zhou, X., Wei, Z., Guo, G.: Energy-based supervised hashing for multimorbidity image retrieval. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 205–214. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_20
9. Johnson, A.E., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**(1), 317 (2019)
10. Khan, A.I., Shah, J.L., Bhat, M.M.: Coronet: a deep neural network for detection and diagnosis of Covid-19 from chest X-ray images. Comput. Methods Programs Biomed. **196**, 105581 (2020)
11. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2014)
12. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The role of imagenet classes in frechet inception distance. arXiv preprint arXiv:2203.06026 (2022)
13. Lim, J.H., Ye, J.C.: Geometric GAN. arXiv preprint arXiv:1705.02894 (2017)
14. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13753–13762 (2021)
15. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Chest X-ray generation and data augmentation for cardiovascular abnormality classification. In: Medical Imaging 2018: Image Processing, vol. 10574, pp. 415–420. SPIE (2018)
16. Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1505–1514 (2019)
17. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
21. Tao, M., Tang, H., Wu, F., Jing, X.Y., Bao, B.K., Xu, C.: DF-GAN: a simple and effective baseline for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16515–16525 (2022)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
23. Xu, T., et al.: AttnGAN: fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1316–1324 (2018)
24. Yang, X., Gireesh, N., Xing, E., Xie, P.: XrayGAN: consistency-preserving generation of X-ray images from radiology reports. arXiv preprint arXiv:2006.10552 (2020)
25. Zhang, H., et al.: StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915 (2017)

26. Zhang, Y., Chen, Q., Yang, Z., Lin, H., Lu, Z.: BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data **6**(1), 52 (2019)
27. Zhu, M., Pan, P., Chen, W., Yang, Y.: DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5802–5810 (2019)