A GAN-BASED HYBRID DATA AUGMENTATION FRAMEWORK ON CHEST X-RAY IMAGES AND REPORTS

by

Hasan Berat Özfidan

B.Sc., Computer Engineering, Istanbul Technical University, 2018

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2022

ACKNOWLEDGEMENTS

Firstly, I would like to offer my special thanks to my advisors Assoc. Prof. Arzucan Özgür and Pınar Yanardağ Delul, Ph.D for their ultimate support during my research, for their motivation, patience and great knowledge. Their guidance supported me throughout the research and the thesis especially during the outbreak of COVID-19 pandemic. I also want to thank my previous advisor Prof. Arda Yurdakul for helping me to start my graduate study.

I also would like to thank my colleagues in Aselsan for helping me to balance the work and academic workload. I want to thank my team leader Furkan Duruk and my colleagues Cevahir Çığla, Burak Özkalaycı, Ömercan Özdemir, Elvan Duruk, Mehmet Emin Orhan, İdris Murat Üzümcü, Gürcan Adalı, Hekim Kahraman, Hakan Demir, Turan Doğan, Oğuzcan Sezgin, İsmail Koton, Onur Balık, Mustafa Karaipek, Erkin Yenigün. I am also thankful to my manager Fatih Özsoy for his inspiring pep talks. Last but not least, I would like to express my gratitude to thank my director Mehmet Kapçak for making this research possible.

I would also like to express my gratefulness to Aselsan Inc. for allowing me to use their high-performance computation resources and graduate study permissions during my employment.

Last of all, I want to express my thanks to Dr. Aycan Alkan for sharing her immense knowledge in the field of Pulmonology.

ABSTRACT

A GAN-BASED HYBRID DATA AUGMENTATION FRAMEWORK ON CHEST X-RAY IMAGES AND REPORTS

Classical data augmentation techniques are widely used by many image classification applications in the absence of adequate training data. These data augmentation techniques consists of but not limited with reflection, random cropping, re-scaling existing images and transformations. These techniques are widely used in practice during training classifiers with extended versions of real-world datasets. Increasing dataset size with realistic synthetic data allows us to improve the classification accuracy by making use of additional realistic variety. With the great representational power of GANs, learning the distribution of real data with a consistent level of variety allows us to generate samples with nearly-unobserved discriminative features. In our approach we used the aforementioned generative capability of GANs by utilizing state of the art GAN augmentation framework titled as StyleGAN2-ADA. After the training SytleGAN2-ADA in class conditional setting, we extended the dataset with different numbers of additional generated samples in order to observe the correlation of accuracy and augmentation strength. We extended our approach by using StyleCLIP to experiment disentangled feature augmentations which is a novel approach in the field of GAN augmentation. To make use of StyleCLIP more efficiently, we fine-tuned CLIP with X-ray images and modified entities which are extracted from corresponding medical reports. We used the DeepAUC framework which is proven to be efficient for multi-disease labelled X-ray classification tasks to test the performance of the GAN augmentation. In our approach, we observed that the classification accuracies were improved compared to without text-manipulated GAN augmented setting.

ÖZET

GÖĞÜS X-RAY GÖRÜNTÜLERİ VE RAPORLARI ÜZERİNE GAN TABANLI HİBRİT VERİ GÜÇLENDİRME YÖNTEMİ

Klasik veri artırma teknikleri, yeterli eğitim verisinin olmadığı birçok görüntü sınıflandırma uygulaması tarafından yaygın olarak kullanılmaktadır. Bu veri artırma teknikleri, yansıtma, rastgele kırpma, mevcut görüntülerin yeniden ölçeklenmesi ve dönüşümlerden oluşur. Bu teknikler yardımıyla gerçek veri kümelerinin artırılmış halinin sınıflandırıcıların eğitimi sırasında kullanılması günümüzde popülerdir. Gerçekçi sentetik verilerle veri kümesi boyutunu artırmak, veri kümesine yeni ve gerçekçi varyasyonlar katılması vesilesiyle bu veri kümesi üzerinde eğitilen sınıflandırıcıların doğruluğunu artırmamızı sağlamaktadır. GAN'ların temsil gücü ile birlikte gerçek verilerin dağılımını tutarlı bir çeşitlilik düzeyiyle öğrenmesi, neredeyse gözlemlenmeyen ayırt edici özelliklere sahip örnekler oluşturmamızı sağlar. Yaklaşımımızda, StyleGAN2-ADA şeklinde adlandırılan GAN veri artırma yöntemini kullanarak GAN'ların yukarıda bahsedilen yaratıcı yeteneklerinden faydalandık. Sınıf koşullu SytleGAN2ADA eğitiminden sonra, başarım ve veri artırma miktarı arasındaki korelasyonu gözlemlemek için veri kümesini farklı miktarlarda oluşturulan ek örneklerle genişlettik. Ayrıca GAN vasıtasıyla veri kümelerini güçlendirme alanında yeni bir yaklaşım olan ayrıştırılmış özelliklere yönelik veri artırma yöntemini denemek için StyleCLIP'i kullandık. StyleCLIP'ten daha iyi faydalanabilmek amacıyla CLIP modelini X-ray görüntüleri ve raporlardan ayıklanmış anahtar cümleciklerle tekrar eğittik. GAN veri artırma yöntemlerinin performansını test etmek için CheXpert yarışması birincisi olan DeepAUC yöntemini kullandık. Yaklaşımımızda sınıflandırma başarımlarının GAN veri artırma yöntemlerinin kullanılmadığı durumlara göre daha yüksek olduğunu gözlemledik.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . viii
LIST OF TABLES . . . x
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiii
1. INTRODUCTION . . . 1
1.1. Motivation . . . 2
1.2. Problem Statement . . . 3
1.3. Summary of Contributions . . . 4
2. BACKGROUND . . . 6
2.1. Introduction to GANs . . . 6
2.2. GAN Frameworks . . . 9
2.3. GAN Augmentation . . . 13
3. RELATED WORK . . . 16
3.1. Medical Image GAN Augmentation . . . 16
3.2. Latent Space Image Manipulation on GANs . . . 19
3.3. Medical Latent Space Manipulations on GANs . . . 22
3.4. Visual-Language Joint Representations . . . 23
3.5. CLIP: Connecting Text and Images . . . 24
4. DATASETS . . . 26
4.1. CheXpert . . . 26
4.2. MIMIC-CXR . . . 27
5. METHOD . . . 29
5.1. Synthetic Medical Data Augmentation . . . 29
5.2. Text-Driven Medical Image Manipulated Augmentation . . . 30
6. EXPERIMENTS . . . 33

6.1. Evaluation Metrics 33
6.1.1. Frechet Inception Distance 33
6.1.2. ROC-AUC 34
6.1.3. PR-AUC 35
6.2. Synthetic Medical Data Augmentation 37
6.3. Text-Driven Latent Space Manipulation 39
7. CONCLUSION 55
REFERENCES 56
APPENDIX A: LEGAL NOTICE 67

LIST OF FIGURES

Figure 1.1. The visualization of the dataset types. 4

Figure 2.1. The concept of generative adversarial networks. 7

Figure 2.2. The concept of GAN augmentation. 14

Figure 3.1. Sample text-based latent space manipulation by StyleCLIP. 22

Figure 3.2. Contrastive Pre-training visualization for a batch size of 5. 24

Figure 4.1. The aim of the CheXpert classification task [5] is to compute the probability scores for each candidate diagnosis. 26

Figure 4.2. The graph for diagnosed disease count per image in the validation set of CheXpert. 27

Figure 4.3. The pie-chart for diagnosed diseases in the validation set of CheXpert. 28

Figure 5.1. A sample dependency parser output of a medical report. 31

Figure 6.1. Receiver Operation Characteristic Curve graph for an example binary classifier. 35

Figure 6.2. Precision Recall Curve graph for an example binary classifier. . . . 36

Figure 6.3. The class conditioned generated sample for cardiomegaly. 37

Figure 6.4. Experiment for neutral text normal and target text lung volumes. 40

Figure 6.5. Experiment for neutral text normal and target text cardiomegaly. 40

Figure 6.6. Original, inverted and manipulated images with positive and negative directions generated by StyleCLIP with fine-tuned CLIP model. 46

LIST OF TABLES

Table 6.1. The results of classification experiments with PCAM. 38

Table 6.2. GAN augmentation experiment results in terms of ROC-AUC. . . 39

Table 6.3. The classification results of style manipulations for original CLIP model for 100 samples. 42

Table 6.4. The classification results of style manipulations for fine-tuned CLIP for 100 samples. 42

Table 6.5. The classification results of style manipulations for fine-tuned CLIP for 1000 samples. 43

Table 6.6. Classifier training accuracies with respect to bare dataset, GAN augmented and style manipulation based GAN augmentation. . . . 43

Table 6.7. Classification results on only synthetic style manipulated data in terms of ROC-AUC score. 47

Table 6.8. Classification results on only synthetic style manipulated data in terms of PR-AUC score. 49

Table 6.9. Classification results of fine-tuning strategies on augmented dataset in terms of ROC-AUC score. Each line was obtained when the mean ROC-AUC score is maximum. 50

Table 6.10. Classification results of fine-tuning strategies on augmented dataset in terms of PR-AUC score. Each line was obtained when the mean ROC-AUC score is maximum. 51

Table 6.11. Classification results of fine-tuning strategies on style manipulated GAN augmented dataset in terms of ROC-AUC score. Each line was obtained when the mean PR-AUC score is maximum. 52

Table 6.12. Classification results of fine-tuning strategies on style manipulated GAN augmented dataset in terms of PR-AUC score. Each line was obtained when the mean PR-AUC score is maximum. 53

LIST OF SYMBOLS

$\mathbb{E}_{\boldsymbol{x}}$	Expectation function for a variable $\boldsymbol{x}$

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
AUC	Area Under the Curve
CNN	Convolutional Neural Networks
FID	Frechet Inception Distance
GAN	Generative Adverserial Network
NLP	Natural Language Processing
PCA	Principle Component Analysis
ROC	Receiver Operating Characteristic
SOTA	State-of-the-art
VAE	Variational Autoencoder

1. INTRODUCTION

Contemporary developments in CNNs have improved the state of the art results in numerous domains such as agriculture, manufacturing, self-driving cars, surveillance, toll-collection, medical imaging and so on. The main source of this success are the high availability of the datasets and enhanced low-priced computing resources. To develop deep learning based computer vision applications, data collection process must be carefully conducted. However, if the collected data is not sufficient, the computer vision application may not be successful as it is desired. Image segmentation, object detection and image classification is the hearth of the almost every deep learning based computer vision application. These tasks utilizes gigantic layers with huge number of parameters that requires to be optimized. Since the "engine" of computer vision applications gets bigger day by day, the engine requires more and more fuel of "data". However, while the number of datasets is increasing, most of the datasets suffers from lack of representational power which is similar to using low-quality fuel. In our approach we used text-based style manipulated GAN augmentation in order to improve the quality and quantity of the data.

Image classification problem one of the fundamental problems in computer vision. To solve this problem, classifiers learn to frame crucial discriminative features for an object in an image. Beside learning these representative features, classifiers also ignore the unimportant parts of an image such as noise, bare background, orientation of an object or size of an object. Most of the widely used deep learning based image classification models are fed with an input image and pass these information throughout several convolutional filters and pooling layers. In the one hand, convolutional filters enable computer vision models to extract discriminative features. On the other hand, pooling layers compresses the data in order to reduce the size of information without hurting representation power. By employing these convolutional filters and pooling layers in a structure, in the end it yields feature maps. All these feature maps are then flattened into a single tensor which is input of the later-coming fully connected layers.

These layers outputs the probability distribution of class scores. In order to predict a class for an image, the index of the highest value among these classes is selected.

Classical data augmentation techniques are widely used by many image classification applications in the absence of adequate training data. These data augmentation techniques consists of but not limited with reflection, random cropping, re-scaling existing images and transformations. These techniques are widely used in practice during training classifiers with extended versions of real-world datasets. Increasing dataset size with realistic synthetic data allows us to improve the classification accuracy by making use of additional realistic variety. With the great representational power of GANs, learning the distribution of real data with a consistent level of variety allows us to generate samples with nearly-unobserved discriminative features. In our approach we used the aforementioned generative capability of GANs by utilizing state of the art GAN augmentation framework titled as StyleGAN2-ADA [1]. After the training SytleGAN2-ADA in class conditional setting, we extended the dataset with different numbers of additional generated samples in order to observe the correlation of accuracy and augmentation strength. We extended our approach by using StyleCLIP [2] to experiment disentangled feature augmentations which is a novel approach in the field of GAN augmentation. To make use of StyleCLIP more efficiently, we fine-tuned CLIP with X-ray images and modified entities which are extracted from corresponding medical reports. We used the Probabilistic Class Activation Maps framework (PCAM) [3] which is proven to be efficient for multi-disease labelled X-ray classification tasks to test the performance of the GAN augmentation. In our approach, we observed that the classification accuracies were improved compared to without GAN augmented setting.

1.1. Motivation

Data is the new "petrol" thanks to the advanced data processing approaches. However, collection of expert-annotated data is one of the bottlenecks in the field of deep learning. Gathering labeled data for object recognition tasks is rather straightforward compared to collecting data for medical decision support applications. Data

collection for medical downstream tasks require expert examinations and cross validations. On the contrary, object recognition tasks do not require further expert knowledge. In practise, even a children can annotate and compile such datasets. Yet, we require board-certified expert medical doctors with years of experience for annotating and creating such datasets. With our approach, we aimed to decrease the need of expert knowledge for medical downstream tasks. Moreover, creating such datasets is not the only challenge since some of the annotations include personal information. In order to make the dataset public, it must be de-indentified. In other words, the annotations should be free of any kind of sensitive personal information. Since the generated data with our approach does not belong to any real person. Our approach relaxes the de-identification processes. This facilitates the contribution to public datasets. Our approach is also able to reveal development phases of the medical conditions with interpolation videos. This media could be used as educational material for medical students.

1.2. Problem Statement

Image classification using deep learning techniques requires large number of image samples in order to train accurate and precise classifiers. However, training classifiers with small and unbalanced datasets such as medical datasets is a challenging task. According to Sampath *et al.* [4] image classifiers mostly designed for performing well with balanced datasets. However, they stated that most of the real-world datasets suffers from imbalance of observed classes. According to their research, the class imbalance problem is well known for image classification tasks. They indicated several different categories for imbalance and size problems of real-world datasets. According to their description, the ideal dataset should include enough amount of data for each class. If a dataset has sufficient amount of data yet the data is not distributed for each class evenly, they categorized that kind of datasets as uneven datasets. If the data is evenly distributed within a dataset yet the total number of samples are not sufficient, they categorized this kind of datasets as tiny datasets. The final category in their work is absolute rare datasets which neither has sufficient data nor inter-class balanced sam-

ples. In our research, we worked on chest X-ray dataset titled as CheXpert [5] which could be considered as absolute rare dataset since the sample counts for each class is extremely uneven distributed. In Figure 1.1, the properties of absolute rare datasets is shown. To compensate the downsides of the dataset, we employed GAN augmentation techniques. Thanks to GANs, it is possible to create realistic samples and augment datasets to train better performing classifiers.

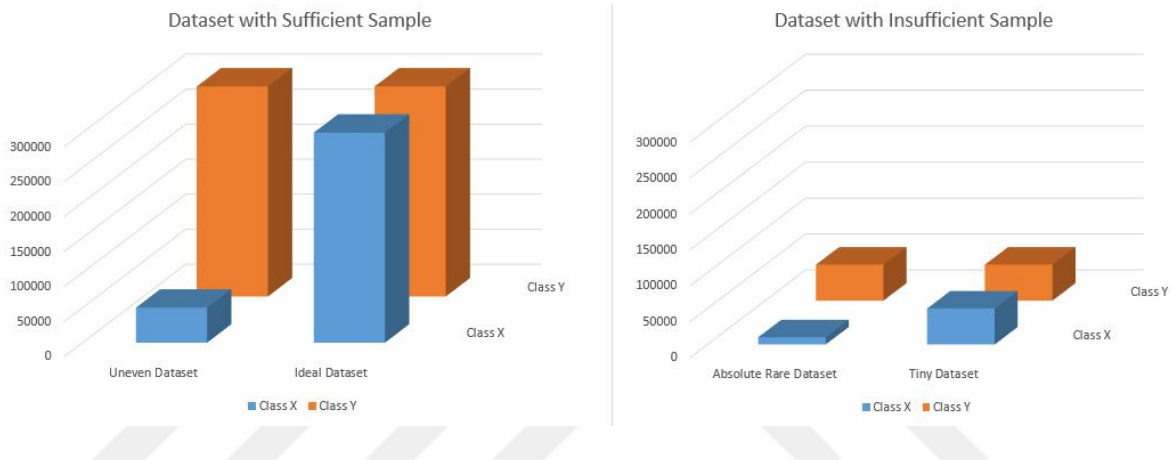

Figure 1.1. The visualization of the dataset types.

1.3. Summary of Contributions

In this work, we presented several contributions in the field of medical GAN augmentation;

- synthetic additional data generation by using StyleGAN2-ADA,
- rule-based algorithm for information extraction from X-ray reports,
- evaluating CLIP by fine-tuning with different text extraction methods for medical domain adaptation,
- using text-based latent space manipulations for data augmentation.

For our best knowledge, we are the first who uses StyleGAN2-ADA to generate additional data for the CheXpert dataset. After data generation in class conditional setting, we trained several classifiers to observe the effect of GAN augmentation to the

accuracies of the classifiers. The results showed that the performance of the classifiers were improved with the help of additional GAN generated samples.

Although CLIP is a very powerful zero-shot classifier, we believed that it has nearly no chest X-ray data in its original training set since the style manipulations with original CLIP model was not consistent. Therefore, we fine-tuned original CLIP model to leverage CLIP's text-image embedding abilities in our problem domain. In our case, since it is not possible to feed all the report data due to the context length constraint of CLIP, we wanted to develop an algorithm which extracts important portions of the reports. To manage this, we used scispaCy models which are trained on large corpus of biomedical text. We extracted entities and entity relations with scispaCy models and build our logic on the top of these outputs to create more machine-friendly reports that do not exceed the size of the original CLIP's context length. By using the fine-tuned CLIP, we experimented with StyleCLIP to see if it is possible to augment the dataset with style manipulations. Our preliminary results showed that it is possible to control disease specific attributes to increase the classification score for a particular disease.

2. BACKGROUND

This chapter covers the background information of our work. For the beginning, we covered the basic concepts of Generative Adverserial Networks (GANs). Then, we included advanced GAN approaches. In the end, we explained GAN augmentation which is one of the key concepts of our research.

2.1. Introduction to GANs

There are different types of approaches when it comes to capture the probability distribution for a given data set. Once the probability distribution for a given data set is captured successfully, it enables us to populate realistic data samples by simply drawing from the captured probability distribution. These approaches mainly leverage statistics, probability theory, linear algebra, multivariate calculus and optimization theory. By utilizing the combinations of these mathematical tools, some popular generative models had been forged. Some of the examples from these generative models could be given as Boltzmann Machine [6], Variational Autoencoders [7], Bayesian Networks [8] and Generative Adverserial Networks [9].

Deep Generative Modeling is an unsupervised learning method in deep learning that includes auto-discovery of the consistencies or patterns in input samples so that the model can be utilized to create new samples that might have been drawn from the original data. However, the main area of this research is Deep Generative Modeling with a focus on Generative Adversarial Networks.

Generative Adversarial Networks were firstly proposed by Ian Goodfellow *et al.* in 2014 [9]. The proposed model consists of two different deep neural networks which are the generator and the discriminator. On the one hand, the goal of the discriminator is to classify whether the output of the generator is fake or real. On the other hand, the objective of the generator is to deceive the discriminator network by generating

realistic fake outputs as shown in Figure 2.1. During the training, both of the networks learn internal representations to master their specific tasks by playing a mini-max game where one is trying to outperform the other. In theory, at the end of this adversarial game, they shall reach an equilibria point which leads the outputs of the generator to be indiscriminable to the actual samples.

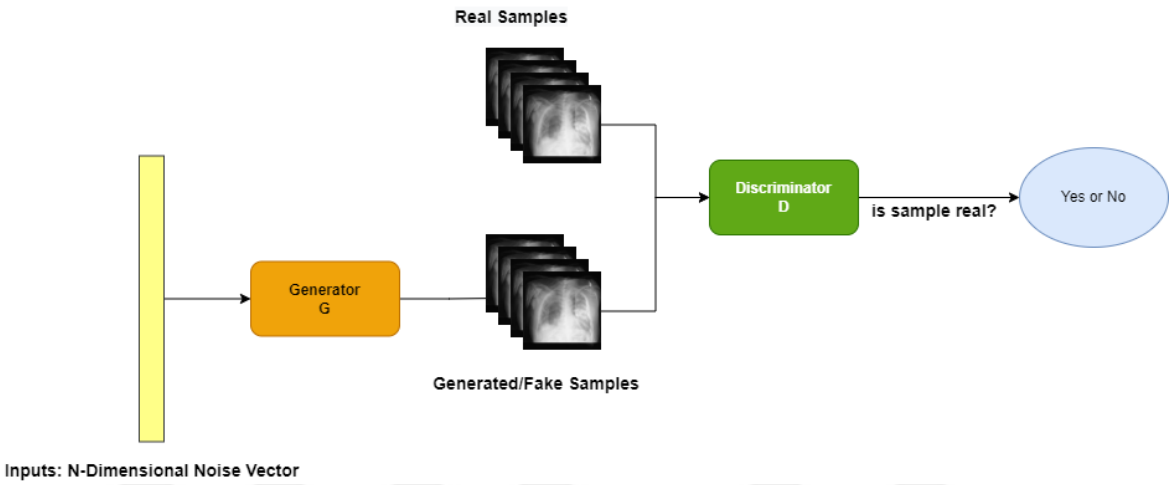

Figure 2.1. The concept of generative adversarial networks.

The generator network takes a fixed-size arbitrary one dimensional tensor as input and creates a sample in the input domain. The tensor is usually drawn a Gaussian distribution, and the tensor is utilized to trigger this creative operation. After the training of GANs, points in this vector space shall represent the points in the input data domain, modeling a squeezed representation of the input data distribution. This vector spaces named as latent or hidden space. Hidden space consists of hidden variables or latent variables which are beyond our observation space. That is the reason why it is also called "hidden" space. The discriminator network is a simple binary classification model and the aim of this network is to determine whether the input data is real or fake. To this end, the discriminator network is fed with real and generated samples. The ratio of real and generated samples used for feeding the discriminator network might vary depending on the training strategy. However, using same number of generated and real data for feeding the discriminator is widely used in most of the approaches.

In the training phase, the discriminator is fed with both real and generated samples. For each real and generated samples, the discriminator assigns probabilities in between **1** and **0**. The assigned probability is the degree of reality for a given input from the perspective of the discriminator. Therefore, realistic samples supposed to be assigned to higher probabilities which are just about 1. Contrariwise, if the input samples are considered as fake the samples are assigned to the probabilities which are close to 0. The discriminator tries to minimize the distance between actual labels and predicted labels. To do so, the discriminator needs to maximize the classification performance. Nevertheless, in the generator's training, the generator is fed with random noise and expected to create such realistic outputs that confuse the discriminator. By confusing the discriminator, the generator tries to maximize the distance between actual labels and predicted labels and that leads the classification performance of the discriminator to decrease.

To comprehend GANs training completely, minimax GAN loss should be examined. For

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}\,(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))], \tag{2.1}$$

the discriminator tries to maximize both $\log D(\boldsymbol{x})$ and $\log(1 - D(G(\boldsymbol{z})))$, which also maximizes the classification accuracy of the discriminator. Conversely, the generator aims to minimize $\log(1 - D(G(\boldsymbol{z})))$ by forcing the discriminator to misclassify its outputs.

Once the GANs training is completed, it enables us to generate almost anything. There are high variety of applications which use the creativity of GANs. Alqahtani1 *et al.* [10] categorizes these applications in five sections which are image based applications, domain adaptation, sequential data based applications, improving classification and recognition, miscellaneous applications. Image based applications includes generation of high quality images, image inpainting, super-resolution, person re-identification, object detection, video prediction and generation, facial attribute manipulation, anime character generation, image to image translation, text to image translation, face aging,

human pose estimation, de-occlusion and image blending. Sequential data based applications consists of music and speech generation. Miscellaneous applications covers drug discovery and molecule development in oncology.

2.2. GAN Frameworks

Since GANs firsty proposed by Goodfellow *et al.* [9] a new era in deep learning has been emerged. According to Google Scholar, "Generative Adverserial Nets" [9] paper recieved over 25000 citations until the begining of 2021. The overall citation count of Ian Goodfellow's academic papers which are mostly related to GANs is over 124.000. A simple comparison can describe the real meaning of the aforementioned citation counts. Albert Einstein who considered to be one of the exceptional scientist throughout the human history received nearly 136.000 citations until the end of 2020. Charles Robert Darwin, the pioneer of "The Theory of Evolution", also received nearly 170.000 citations until 2021. These numbers show that even for 7 years of period GANs are influential on not only deep learning history but also entire science history. From another point of view, the growing numbers of citations with respect to GANs indicate that it is a highly active research field. In the literature, numerous variants of GANs have been proposed. Some of the fundamental GANs variants could be given as CGAN [11], DCGAN [12], LapGAN [13], InfoGAN [14], EBGAN [15], WGAN [16] and so on.

Mirza and Osindero [11] proposed CGAN in 2014. Conditional GAN (CGAN) is basically the conditional version of vanilla GANs. While training both of the generator and the discriminator, there is also an additional condition parameter is included during the learning process. The conditional parameter helps CGAN to model different output data domains in only a pair of neural architectures which are the conditional generator and the conditional discriminator. In other words, with this approach we are able to compress many GANs into just one CGAN with a single controlling parameter. By applying their method, it becomes possible to overcome scalability problem of output classes. Moreover, their approach also tackles the problem of one-to-one mappings

from input to output in the context of GANs. They experimented with two different datasets which are MNIST and Flickr image dataset with user tags. They managed to generate realistic class conditional outputs for each digit class after the training with MNIST. For the second dataset, they adopted their approach for generating user tags for a given input image. Once the training is done, their class conditional model was able to create semantically meaningful tags with respect to input images.

Radford *et al.* [12] addressed the issue that unsupervised learning with CNNs has seen little adaptation compared to supervised learning in computer vision applications. They proposed Deep Convolutional Generative Adverserial Networks (DCGANs) framework to lure attention on unsupervised learning applications. In their approach, they described a set of constraints for Convolutional GANs' architectural topology in order to make the training process stable without scalability issues. To do so they replaced all kinds of pooling layers with stride convolutions. They also used Batch Normalization for both the generator and the discriminator. For the generator, they used only ReLU for activations. For the discriminator, they chose Leaky-ReLU as squashing function. In their work, they compared the trained discriminators with other unsupervised classifiers and showed the performance likelihood between them. Moreover, they visualized the learned filters by generator and they explored the latent space with the help of vector arithmetics. Denton *et al.* [13] described a sequential generative model which is able to produce high quality and natural looking images. With LapGAN, they brought convolutional networks, GANs and Laplacian pyramid framework together. They trained GANs layer by layer within the Laplacian pyramid structure in order to encapsulate particular levels of details in each layer. In the generation phase, they consubstantiate generative models sequentially to build a single high grade realistic output sample. In their approach, they used LSUN and MNIST, STL and CIFAR10 datasets. For quantitative evaluation, they applied Parzen window based log-likelihood estimates in order to compare their method with original GANs. For qualitative evaluation, they developed an application which shows an image to the participants and ask whether the shown image is real or fake. They analyzed the results after obtaining the output of the evaluations.

Chen *et al.* [14] introduced a framework named Information GAN (InfoGAN) which extends GANs to the direction of information theory. In their approach, InfoGAN model is capable of learning interpretable disentangled representations of input data without any supervision. The objective function of InfoGAN is to maximize the variational mutual information between a small set of hidden variables and the observed samples. With the help of lower bound thresholding of mutual information objective, their model managed to extract writing style from digit forms on the MNIST dataset. Their model is also able to disentangle some visiual concepts which are hair styles, absence or presence of eyeglasses and emotions. Their experiments show that their approach archived to disentangle pose, lightning, elevation, rotation and width on 3D objects with the aid of latent vector manipulation. Their method consumes negligible additional computational resources compared to vanilla GANs. Therefore, authors stated that their method does not bring any extra computational cost during the training procedure.

Energy-Based GAN (EBGAN) was proposed by Zhao *et al.* [15] in 2016. EBGAN could be described as the naive combination of auto encoders and GANs. Unlike conventional GANs, EBGAN utilizes the discriminator as an auto encoder with energy function where the real samples are assigned to low energy values and the generated samples are assigned to high energy values. Therefore, the objective of the generator is the minimize the energy value of generated samples and maximize the energy value of actual samples. With EBGANs, authors managed to stabilize the training process and scale the generation of high-quality images. They compared the convergence patterns between conventional GANs and EBGAN with the aid of a variant of the Inseption Score (IS) . They also analyzed Ladder Network bottom layer cost of EBGAN and two of the Ladder Network variants. In their work they also observed that when the generator in their framework was distant from convergence, the perfomance and the quality of the gradients were increased.

Arjovsky *et al.* [16] introduced a GANs framework named Wasserstein GAN (WGAN). In the contrast to conventional GANs, WGAN includes a novel loss function

derived from Earth-Mover or Wasserstein distance. With this approach, they were able to make the training process more stable and partially overcome the problem of mode collapse. They also tackled vanishing gradient problem. Thanks to the expressive learning curves they provided, hyperparameter tuning and debugging became less effort consuming.

Boundary equilibrium GAN (BEGAN) was described by Berthelot *et al.* [17]. In their proposed approach, they introduced a novel equilibrium enforcing method combined with Wasserstein distance based cost function and auto encoder based GANs. The main purpose of their proposed equilibrium enforcing concept is to balance the discriminator and the generator at training time. Their method also offered a controlling technique for the trade-off between quality and diversity of the generated samples. Moreover, they provided a global convergence measure with the help of the equilibrium concept. With the global convergence measure, it became possible to understand whether the model has collapsed or the model reached its final stage.

Karras *et al.* [18] proposed a different framework named as Progressively-Growing GAN (PGGAN) which is able to grow both the discriminator and the generator stage by stage. The model starts to training with low resolution images and progressively increases the depth of both the discriminator and the generator with additional layers. At the end of the training, the model becomes to be able to generate high resolution images. This approach offers high resolution (1024×1024) and high quality image generation along with optimized training stability and decreased training time.

Generation of high resolution images with high fidelity is one of the challenging tasks in the context of generative modeling. In order to overcome this challenge, Brock *et al.* [19] proposed a computationally expensive yet effective framework called BigGAN. Compared to prior state of the art frameworks, the neural networks in this framework have nearly four times as many parameters. However, they were confronted by stability issues due to the huge number of parameters and architectural changes. They introduced a novel approach named as "Truncation Trick" in order to stabilize

the training. Compared to prior works, they also improved Inception Score (IS) from 52 to 166 and Frechet Inception Distance (FID) from 19 to 7 for the down-scaled version (128×128) of ImageNet dataset. Moreover, they also presented image interpolation application which could be simply described as predicting the middle image between to consecutive images.

Although generating high quality images with high resolution is an immense challenge, disentanglement of the style from an original image in a controlled manner is another essential challenge. To tackle that issue, Karras *et al.* [20] presented another framework titled as StyleGAN which is a style-based generator architecture for GANs. In fact, this framework could be considered as improved version of PGGAN [18] with additional style disentanglement and stochastic variation features. Their approach enables unsupervised controllable separation of fine-grained features without damaging the quality and the fidelity of the generated images. One of the significant contributions of this approach is the utilization of the mapping network to create intermediate latent vectors along with Adaptive Instrance Normalization (AdaIN) which provides the control over the generator during the style transfer process. For quantifying the style disentanglement, they described two novel metrics which are perceptual path length and linear separability. After this work, they proposed an improved version of StyleGAN called StyleGAN2 [21] after discovering the fact that water droplet-like artifacts was caused by instance normalization. They redesigned the generator normalization and improved progressive growing. In the end, StyleGAN2 outperformed the prior state of the art results in the field of unconditional image modeling.

2.3. GAN Augmentation

GAN Augmentation is a relatively new approach for augmenting training data. Before GAN augmentation there were some classical augmentation techniques which includes yet not limited with random cropping, translation, transformation, reflection, change of orientation, re-sampling and re-scaling. The training data is extended via these operations to improve the classification accuracy . However, using GANs for

generating additional training data is becoming active area of research in recent years due to the facts that it does not require hand-crafted techniques and it has impressive representational power of input data. The concept of GAN augmentation could be found in Figure 2.2.

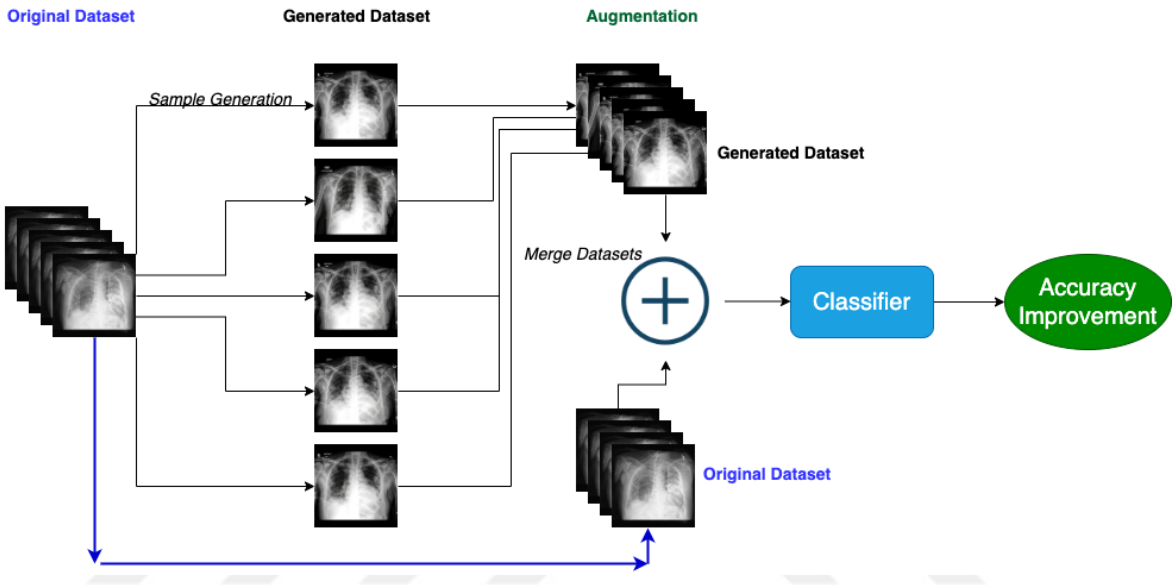

Figure 2.2. The concept of GAN augmentation.

Bowles *et al.* [22] proposed that the generated samples could be used for populating the training data. To do so, they used the PGGAN [18] architecture for modeling the input data distribution. They combined the classical data augmentation techniques with GANs augmentation in their experiments. Their experiments showed that using both of these augmentation techniques are beneficial for improving classification precision. They also compared the classical augmentation techniques with GANs augmentation. They conducted their experiments on different datasets in order to measure the generalization performance of synthetic data augmentation strategy. Moreover, they applied different extends of augmentations to observe the effect of augmentation strength.

Mariani *et al.* [23] created a framework named as Balancing GAN (BAGAN) which aims to regularize unbalanced datasets in order to improve deep learning based image vision tasks. They utilized adversarial learning for both majority and minority classes. The model learned discriminative features from majority classes and extends the minority class samples by effectively utilizing the learned features. Their architecture includes an encoder module of an auto-encoder for improving latent space representation of class specific features. Their experiments showed that BAGAN generated images outperformed other state of the art adversarial networks in terms of fidelity in the presence of imbalanced training data.

AugGAN is an GANs augmentation framework proposed by Huang *et al.* [24]. With this framework, they improved the results of classical deep learning based image-to-image translation methods with adversarial training. The network consists of encoders, generators, disciminators and parsing networks. They observed that, training Faster-RCNN and YOLO models with additional augmented data generated by AugGAN was improved the object detection performance.

Lim *et al.* [25] introduced Doping framework in order to generate realistic samples for extending training datasets. Unlike other approaches, they focused on GANs augmentation for unsupervised anomaly detection task. In their methodology, they oversampled infrequent normal samples which are the main cause of false positives in anomaly detection. However, working with high-dimensional data with multi-model distributions is a challenging task for additional data generation. They reduced the complexity of the datasets by using adverserial autoencoders. After, they collected samples at the edge of the latent data manifolds which populates infrequent normal samples successfully. According to their statement, their technique is the first data augmentation method which deals with unsupervised anomaly detection problem. In their experiments, they observed improvements on several datasets.

3. RELATED WORK

In this chapter, we will cover the literature review related to our approach. Our literature review focuses on Medical Image GAN augmentation, Latent Space Image Manipulation on GANs and Medical Latent Space Manipulations on GANs.

3.1. Medical Image GAN Augmentation

Medical GAN augmentation is an active area of research. There are numerous approaches use GAN data augmentation in medical domain. For instance, Frid-Adar *et al.* [26] collected data from Sheba Medical Center for three specific diseases. They adapted DCGAN, ACGAN and ACGAN discriminator to their experiments in order to explore the effect of GAN augmentation for the classification task. They trained classifiers for three different settings which are augmentations only, with additional real images and with additional generated images. In their experiments, the classifier which trained with additional generated images performed best. Kiyasseh *et al.* [27] augmented time series dataset for their medical diagnosis classification problem. They proposed their own framework called PlethAugment which utilizes three different conditional GAN architectures which are CGAN with data sensitivity, DeLiGan with data sensitivity and MADGAN. According to their experimental results, MADGAN outperformed two of the other conditional GANs. Frid-Adar *et al.* [28] also used GAN augmentation for medical data. The aim of their work is classificaiton of three different liver lesions. Their classification experiments focused on training models with classical data augmentation and additional generated images. For generating class conditional samples, they adapted DCGAN architecture. The results of their experiments showed that compared to classical data augmentation, they observed nearly 7% improvement on sensitivity and 4% improvement on specificity. Similarly, Madani *et al.* [29] also examined effects of classical data augmentation and GAN data augmentation. For GAN augmentation, they used DCGAN architecture. They utilized DCGAN for generating more training samples. They compared the classification accuracies for no augmen-

tation, classical data augmentation and GAN data augmentation. The classification accuracies where 81.93%, 83.21% and 84.19% respectively. Bhattacharya *et al.* [30] aim was to classify three medical conditions from chest X-ray images. They used DCGAN architecture for augmenting training data. They conducted experiments on NIH Chest X-ray dataset. The results of their experiments showed that the CNN classifier trained with augmented dataset performs better than the classifier without any augmentations. Xing at al. [31] also experimented on NIH Chest X-ray dataset. Their approach explores the effect of GAN augmentations for lung disease classification and localization . The authors used StarGAN, Pix2Pix and Pix2Pix-N architectures in their experimental settings. They invited certified radiologists to evaluate their generated samples. Both qualitative and quantitative results showed that generated images from Pix2Pix and Pix2Pix-N are better quailty than StarGAN generated images. Moreover, dataset augmentations with Pix2Pix and Pix2Pix-N performs better at disease localization compared to StarGAN augmention. Kora Venu *et al.* [32] conducted experiments with both classical data augmentation and GAN data augmentation in pneumonia recognition task. Their results showed that data augmentation performs well compared to no augmentation settings. However, DCGAN augmentated classifier performs best in terms of accuracy, recall and F1 score. In this work, Ganesan *et al.* [33] discussed traditional augmentations and GAN based augmentations in the context of pneumonia recognition task. They compared the results of the experiments that cover the effects of classical augmentations and PGGAN. Unlike other related work, their results show that classical augmentations outperform GAN based augmentations in terms of disease classification accuracy. Malygina *et al.* [34] experimented with CycleGAN in order to augment their data for multi-class disease classification problem. In their experiment settings, they genareted opposite class samples from CycleGAN trained with ChestXray14 dataset. Their approach performed well on pneumonia and pleural-thickening classification task compared to non-augmented setting. However, the classification accuracy of fibrosis decreased. Mahapatra [35] proposed a method for domain adaptation for medical image registration. He experimented with NIH ChestXray14 dataset. His method facilitate deep learning models to learn different types of images apart from their learned domain. He utilized convolutional autoencoders and Cycle-

GAN to enable domain adaptations. According to his experiment results, a model trained on one of chest X-ray, brain or retinal MR datasets performs better on other two domains compared to conventional methods. Motamed *et al.* [36] also utilized GAN augmentation in medical setting. Their aim was to examine the effects of GAN augmentations on multi class disease prediction task. They trained RANDGAN and AnoGAN models with Covidx dataset. After training, the results showed that compared to AnoGAN, RANDGAN poses better classification accuracy. Segal *et al.* [37] also explored GAN data augmentation on medical domain. Their approach includes experimentation with PGGAN. They trained PGGAN model with NIH ChestXray14 dataset. They managed to reach 8.02 FID score from their GAN training. For data augmentation, Menon *et al.* [38] applied transfer learning from Kaggle Pneumonia X-Ray dataset to their own approach named MTT-GAN. After employing transfer learning, they trained MTT-GAN with covidchestxray-dataset. They compared binary Covid classification results between several ablations of their method. Khalifa *et al.* [39] used GANs to augment the dataset they used. Their proposed method utilizes deep transfer learning and fine tuning on AlexNet, GoogLeNet, Squeeznet, and Resnet18 architectures to classify pneumonia cases. They used 10% of real samples and 90% of generated samples. The classifier with Resnet18 backbone performed best according to their experimental results. In another academic work, Kovalev and Kazlouski [40] employed DCGAN and PGGAN architectures to create realistic samples. They trained classifier with only real images and with only generated images and compared the classification results. Salehinejad *et al.* [41] argued that class inbalance is one of the possible reasons for low classification accuracies for medical datasets. They balanced the dataset by sampling outputs from DCGANs. The experiments showed that the classifiers trained on balanced medical dataset performs better compared to other classifier trained on inbalanced medical dataset. Chen *et al.* [42]proposed a novel GANs based technique for domain adaptive chest X-ray image segmentation method. Their method transforms the test image to the source domain of the pretrained segmentation model. For preserving pixel level and semantic structural content, they introduced a cycle-consistency loss and a semantic-aware loss respectively. The experiments in their work showed comparable results to the state-of-the-art supervised transfer learning

methods. Terzopoulos *et al.* [43] proposed a framework named MAVENs which is a combination of GANs and VAEs. Their framework consolidates both adversarial learning and variational inference simultaneously. They experimented with several datasets including chest X-ray datasets. Lanfredi *et al.* [44] employed VA-GAN, DeFI-GAN to ADNI and COPD datasets. Their approach indicates deformation fields of medical images and shows evidance of anormalies. Deepshikha and Naman [45] proposed a framework for GANs augmentation named Polarity-GAN. With Polirity-GAN, they offered a solution for class overlap problem in conditional GANs. Their approach compares several augmentation methods which are SMOTE-SVM, Resnet+RF, Resnet5, AC-GAN, BAGAN, WGAN-GP and Polarity-GAN in terms of classification accuracy. Training on CovidX dataset is also covered by their experiments. Choong *et al.* [46] proposed a training pipeline which utilizes ACGAN, PGGAN and transfer learning. They compared results of their pipeline within several ablations of their method. They showed that using only one component from the components of their pipeline is not sufficient for archiving best performance. Research by Sundaram and Hulkund [47] indicated that using GANs for improving the classification performance of underrepresented classes outperforms classical data augmentation techniques. However, they only reported the results in terms of ROC-AUC score which is not sensitive to data imbalance. In our paper, we reported PR-AUC scores along with ROC-AUC scores. We also used conditional StyleGAN2-ADA model for synthetic data generation. However, they used CGAN which is a primitive GAN architecture compared to StyleGAN2-ADA. They also focused on only three rare classes which are not subjects of the CheXpert competition. In our approach, we focused on five classes which are the main subjects of the CheXpert competition. In their approach, they used DenseNet-121 architecture for classification. However, we used DeepAUC framework which is the top solution of CheXpert competition.

3.2. Latent Space Image Manipulation on GANs

Latent Space Manipulation on GANs is a very hot topic since GANs attract deep learning practitioners' attention extensively recently. However, the first spark for

Latent Space Manipulation came from Radford *et al.* [12] by applying vector arithmetic to Z space and showing the first results of latent space manipulation. In their work, they showed it is possible to add glasses or smile to the generated outputs by applying vector arithmetic in Z space. To this end, they averaged the Z values generate smiling woman faces and they also averaged the Z values generate neural woman faces. They subtracted the second average from the first and they added the resulting vector to the averaged Z values generate neutral man faces. After the addition, they fed the generator with the output of the addition and the generator yielded an image with a smiling man. In this approach they were able to disentangle smiling attribute in the latent space. Moreover, they also conducted a series of experiments for finding the face position attribute vector. After finding the attribute vector, they applied the interpolants of the attribute vector. The resulting generated images showed the effect of a face turning left from right gradually.

The latent space was also examined in terms of geometrical representations by Arvanitidis *et al.* [48] and they showed using Riemannian metric improves the probability distribution of the latent space in VAEs. Yet, they stated that it is applicable for all generative models. Bojanowski *et al.* [49] proposed GLO for learning better latent space during the training. Their experiments show smooth interpolation results on CelebA. They also studied 4 principle vectors in Z space and showed that these principle vectors control background brightness, face pose and gender attributes. However, their semantic control over the Z space was highly entangled and their principle vectors was only able to control course attributes of the generated images. In other words, according to their experiments, the principle vector controls the brightness also controls hair style and even gender. Upchurch *et al.* [50] proposed Deep Feature Interpolation for controlling image attributes such as facial hair or age. Their method is depending on linear interpolation on deep convolutional features from pre-trained CNN such as VGG-19 trained on ILSVRC2012. They showed comparable results with respect to AEGAN in terms of controlling semantic attributes.Antipov *et al.* [51] proposed Age-cGAN for generating aged face images without losing face identity.Jahanian *et al.* [52] worked on "steerability" of GANs. They observed that simple walks on la-

tent space achieve as powerful transformations as complex walks for the output space. They also showed that there is a hard limit for each transformation depending on the dataset variability. They observed the transformations show different disentanglement properties with respect to different architectures. By using data augmentation, they showed larger transformation effects. Goetschalckx *et al.* [53] proposed a framework named GANanalyze which is able to control memorability, aesthetic and valance attributes. However, their transformation function is only able to capture the attributes depending on the assessor network therefore it requires additional network for controlling each attribute. Shen *et al.* [54] further investigated the latent space in the field of face generation. They observed that after linear transformations, latent space is able to learn disentangled representations for well-tuned generative models. In their experiments, they showed their approach controls eye-glasses, smile, age, pose and gender attributes of the generated faces without touching other semantic attributes. They also used GAN inversion methods for finding the exact Z value for a particular face image. The inversion method allows semantic latent space manipulation for a real face image. Galatolo *et al.* [55] used CLIP in order to find the optimal Z value for generating text to image and vice verse. They experimented their approach with StyleGAN, BigGAN and GPT2.With the recent developments from OpenAI [56], Patashnik *et al.* [2] proposed a framework called StyleCLIP which enables semantic meaningful latent space manipulation with text guidance. In their framework, they worked on the latents spaces of StyleGAN, StyleGAN2 and StyleGAN2-ADA. They employed and experimented with three different approaches which are text-guided latent optimization, latent residual mapper, mapping a textual input-agnostic data into global direction. In one hand, text-guided latent optimization works on W+ space and takes several minutes to find an optimal style manipulation direction vector for a given text and image pair. On the other hand, Latent residual mapper requires 10 to 12 hours for training the mapper network on a single NVIDIA GTX 1080Ti GPU. Latent residual mapper has the control over several attributes per text for an image. For instance, they were able to generate Angela Merkel's face image with curly and long hair by only inputting "Curly long hair" in the text field. For a given pair of text which defines neutral and target attributes, global direction method exploits the colinearity between CLIP's image em-

bedding space and StyleGAN's style space for the difference of the given text pair. In other words, they found that given a pair of text, the corresponding difference vector in the image embedding space is colinear with the StyleGAN's style space. Collins *et al.* [57] also worked on StyleGAN pre-trained model and observed that the representations learned by the generator are highly disentangled corresponding to the semantic attributes of the generated image. They managed to transfer semantic attributes from a reference image to target image partially. Harkonen *et al.* [58] applied PCA to latent space and feature space of StyleGAN. With the help of the principle vectors, they controlled semantically meaningful attributes of the generated image. A sample figure of image manipulation could be found in Figure 3.1. Their experiments also showed the semantic attributes of the generated image by BigGAN could be manipulated by using principle vectors. Wu *et al.* [59] showed a method for exploring high number of distinct and disentangled style channels. They expanded their method with the help of pre-trained image classifier and small set of sample images.

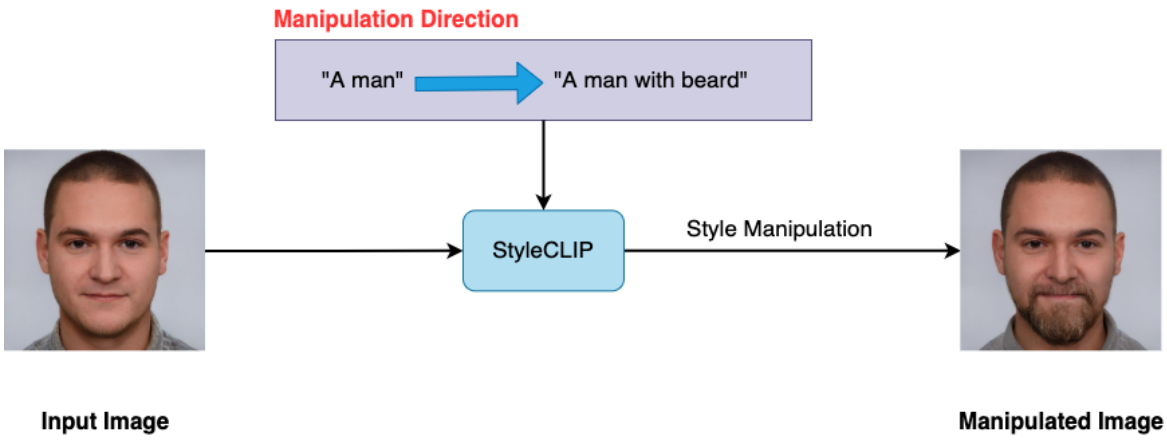

Figure 3.1. Sample text-based latent space manipulation by StyleCLIP.

3.3. Medical Latent Space Manipulations on GANs

Latent space manipulations of GANs is a novel direction of research and it has been studied for just a couple of years. Thus, there are limited studies cover latent space manipulations on GANs in medical approaches. One of the these studies were conducted by Fetty *et al.* [60]. They used StyleGAN with both Computed Tomography

(CT) and T2- weighted Magnetic Resonance (MR) images and explored the latent space walks between modalities. In their experiments, they were able to generate synthetic CT image corresponding to a MR images and vice versa. Yet, they did not further experiment the disentanglement properties of the style transfer. Fernandez Blanco *et al.* [61] worked on histopathological image dataset and produced Grad-cam heatmap to show the features of real and manipulated images match.

3.4. Visual-Language Joint Representations

There are many recent approaches for the joint representations of visual-language domain. Research by Desai and Virtex [62] showed that using a pre-trained VirTex model, it is possible to use the pre-trained model for other downstream tasks such as instance segmentation, object recognition and image classification by re-training the model with a few samples. In other work, Sarıyıldız *et al.* [63] proposed a method named ICMLM which is a proxy task that learns visual representations from text and image pairs. In ICMLM, they combined image and text encoders to match visual and textual embeddings. By using ICMLM, they were successfully predict the masked word for a given text-image pair. Another work [64] also combined language and vision domains within a model named LXMERT and tested their approach on several downstream tasks such as masked language modeling, masked object prediction, and image question answering. They reported that their approach was able to reach SOTA results on GQA and VQA visual question answering datasets. With the emergence of BERT [65], which offers quite powerful language modeling representations, some visual-language approaches employed BERT just like VL-BERT [66]. VL-BERT is built on top of transformer and accepts text and image embeddings as input. In VCR benchmark competition, VL-BERT managed to take the first place on the leadership. A recent approach named as CLIP [56] from OpenAI offers zero-shot image classification with the help of the image-text similarity estimations. To enable zero-shot image classification, OpenAI used nearly 400 million image-text pair collected from web during CLIP's training. It is showed that CLIP was able to predict class labels from various domains without any further training or fine-tuning.

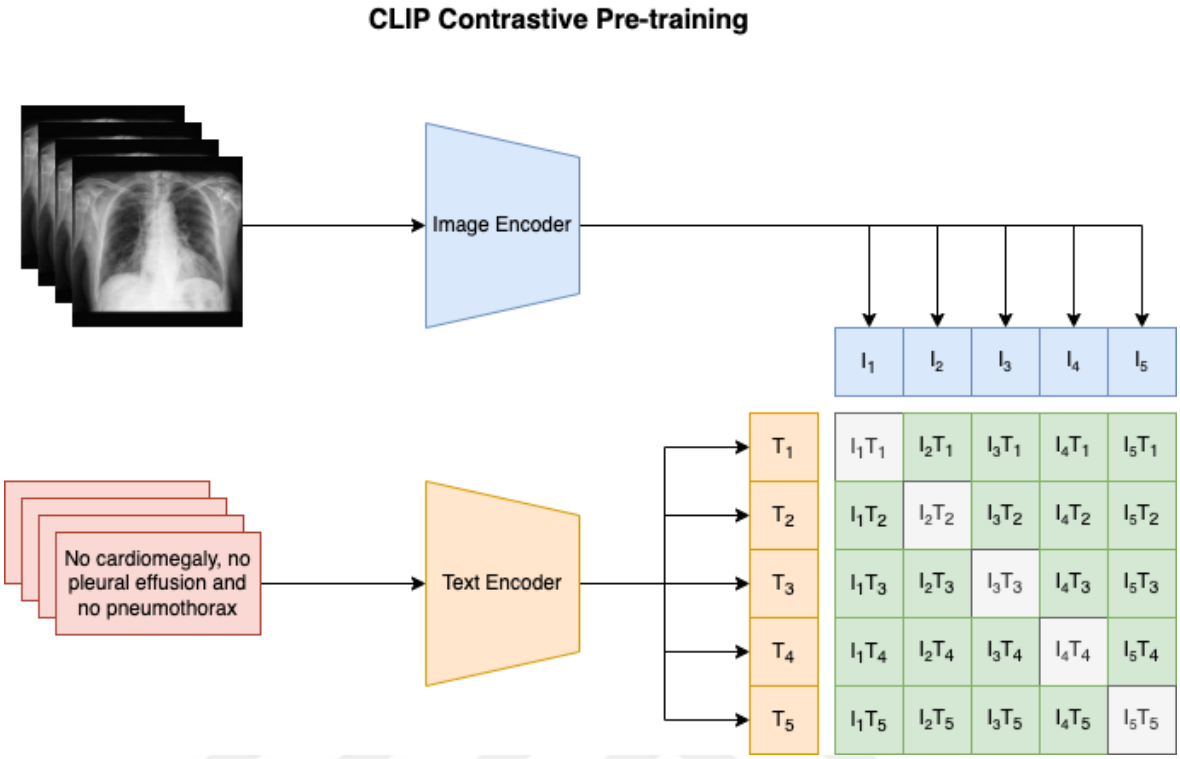

Figure 3.2. Contrastive Pre-training visualization for a batch size of 5.

3.5. CLIP: Connecting Text and Images

Radford *et al.* [56] developed a generic zero-shot approach for bridging the gap between visual and textual domains. For this purpose, they utilized image and text encoder with joint training strategy in order to learn multi-modal embedding space. During training, CLIP uses batches of text-image pairs and for each text it combines the text to all images of the batch. For a batch of M image-text pairs, it yields M accurate pairs and $M^2 - M$ inaccurate pairs. For accurate pairs, the objective function maximizes the cosine similarity of the text-image embeddings. On the contrary, objective function minimizes the cosine similarity of the text-image embeddings for inaccurate pairs. The main objective function optimizes a symmetric cross entropy loss by using the cosine similarity scores of all similarity scores for a given batch. This sort of training strategy is named contrastive pre-training. See Figure 3.2 for visualization of the contrastive pre-training of CLIP.

CLIP consists of two main networks which are image encoder and text encoder. As image encoder, they experimented with modified version of ResNet-50 [67] and Vision Transformer (ViT) [68] with minor differences. As text encoder they employed Transformer [69] with slight architectural modifications. Due to the computational efficiency, they fixed the sequence length of the text encoder to 76. Because of this constraint, we trimmed our text sequences in our experiments. All in all, CLIP combines all of these components and approaches in order to train a single model which is able to perform well on different downstream tasks without any further dataset specific training. The authors tested CLIP with over 30 different computer vision tasks for instance object classification, optical character recognition, action recognition in video footage. It is observed that the performance of the CLIP is mostly comparable to the corresponding complete-supervised baselines.

4. DATASETS

This chapter covers the datasets we used in our work. We used MIMIC-CXR to obtain medical report and X-ray image pairs. We also used CheXpert for measuring the performance of the augmentation techniques employed in this work.

4.1. CheXpert

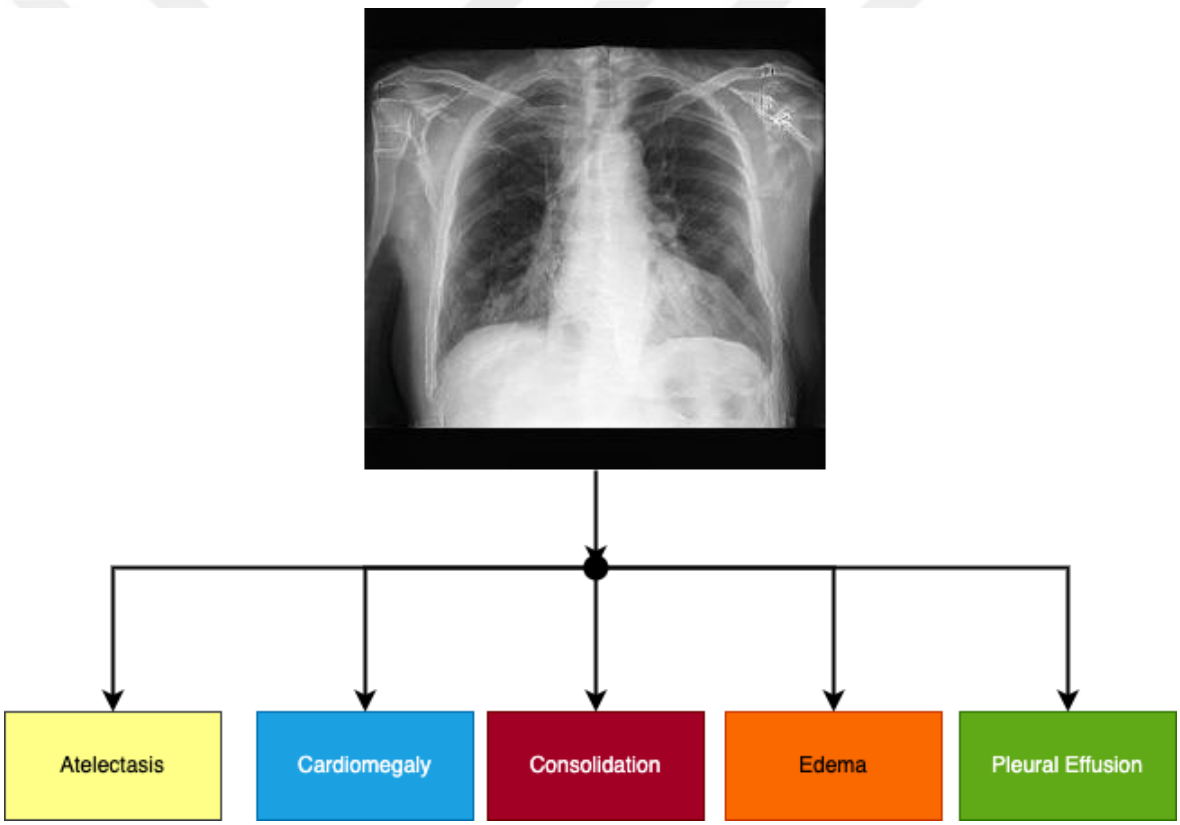

Figure 4.1. The aim of the CheXpert classification task [5] is to compute the probability scores for each candidate diagnosis.

Irvin *et al.* [5] provided a large dataset named CheXpert that includes nearly 225.000 chest radiographs of 65.000 patients. They developed an automatic labeling tool which looks for appearances of 14 conditions in chest X-ray radiology reports corresponding to the X-ray image samples. In the validation set, there are 200 chest X-ray images labeled with corresponding diseases. Unlike the training data, the validation

set was labeled by certified radiologists instead of automatic labeling tool. For each sample in the dataset, there are at least one indicator for each medical condition. In the data collection and label selection process, they used mention extraction, mention classification and mention aggregation techniques. After employing these techniques, they compared their label classification performance with respect to NIH Labeller [70]. Despite the fact that they extracted the labels from the radiology reports, they were unable to share the radiology reports publicly due to the lack of de-identification of the reports according to 25th issue of the their GitHub repository [5] .The overview of CheXpert classification task could be found in Figure 4.1. The analysis of validation set for disease count per image and disease distribution could be found at Figure 4.2 and Figure 4.3 respectively.

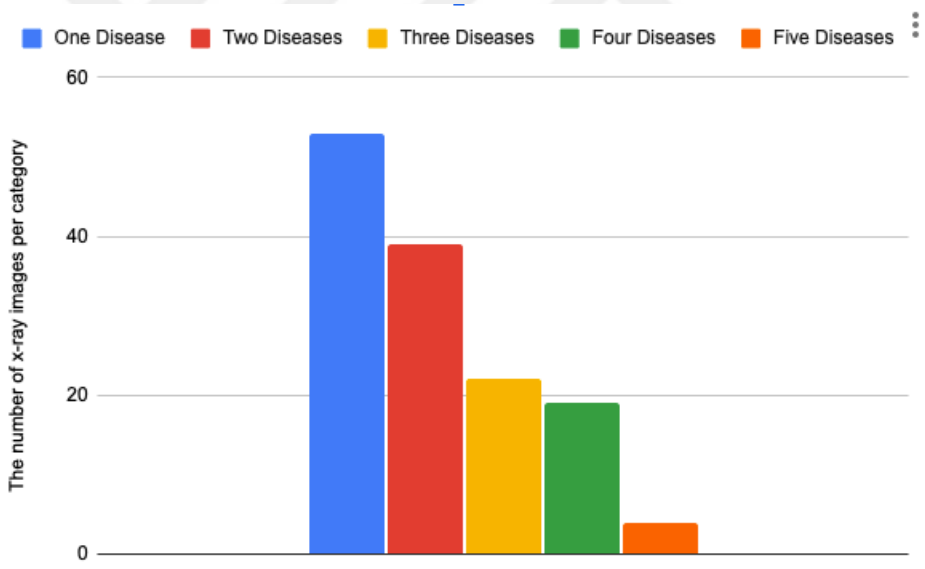

Figure 4.2. The graph for diagnosed disease count per image in the validation set of CheXpert.

4.2. MIMIC-CXR

MIMIC-CXR is a publicly available dataset with 377,110 images and 227,835 corresponding medical studies with respect to these images. The dataset was created by Jonhson *et al.* [71] in collaboration with Beth Israel Deaconess Medical Center in Boston. The dataset was de-identified compliant with the US Health Insurance

Portability and Accountability Act of 1996 (HIPAA) Safe Harbor regulations. For the regulation compliance, they removed Protected health information (PHI) on both radiology images and reports. They used image processing and NLP techniques in order to find the pixel coordinates where PHI burned in the radiology images and they blacked out the bounding boxes including PHI. The PHI removal action could be seen in some of the radiology images as rectangular black censor boxes on the left upper part of the radiology image. They also replaced PHI in the radiology reports with three underscores to protect sensitive information of the patients. In the radiology reports, normal and abnormal findings were placed such as *"lung volumes are normal"* and *"mild enlarged cardiomediastinal silhouette"*. Their aim of creating and making the dataset public was to improve the efficiency of the medical resources especially for the healthcare centers with limited medical resources with the help of the accurate automated analysis of radiology images.

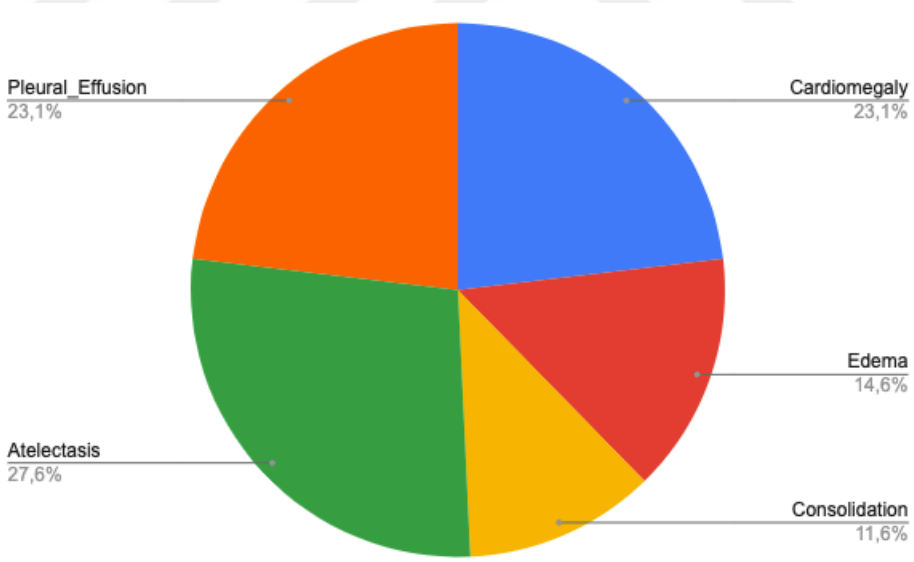

Figure 4.3. The pie-chart for diagnosed diseases in the validation set of CheXpert.

5. METHOD

In this chapter, we will cover the details of our approach. First of all, we will explain how we augment the CheXpert dataset by using GANs. Afterwards, we will cover how text-driven medical image manipulation helps dataset augmentation in our case.

5.1. Synthetic Medical Data Augmentation

We propose to use the synthetic data augmentation pipeline titled as ChestGAN in order to improve the multi-label disease classification performance. Our GAN based augmentation pipeline consists of StyleGAN2-ADA framework [1] to augment CheXpert [5] dataset and Probabilistic-CAM pooling (PCAM) framework [72] for multi-disease classification. To our best knowledge, our approach is the first one which utilized StyleGAN2-ADA framework to augment chest X-ray datasets.

StyleGAN2-ADA [1] is a GAN framework which is designed for generating realistic samples without requiring extensive dataset volume. In this method, the authors proposed to use an adaptive discriminator augmentation technique which stabilizes the training of GANs with tiny datasets. Their GAN augmentation method allows us to generate diverse yet realistic outputs without any augmentation leakage or distortion. According to their experimental results, they improved FID score from 5.594 to 2.42 for CIFAR-10 dataset. Probabilistic-CAM pooling framework [72] is a novel approach which utilizes global pooling operation for indicating the findings of chest diseases with image-level supervision. PCAM pooling utilizes CAM [3] with probabilistic training. Probability maps generated from their approach show sharp and accurate boundaries. In their experiments, PCAM pooling framework improved state of the art results for the classification and localization tasks for the ChestX-ray14 [70] dataset. Their framework also performed best in CheXpert competition with $0,929$ AUC score.

In our approach, we aimed to generate synthetic data samples in order to augment the CheXpert dataset. Unlike traditional data augmentation techniques, GAN augmentation offers both creative and realistic sample generation which enables us to model the data distribution of existing datasets with an conservative extent. By employing this approach, we believed that we were able to create fake samples preserving the distinctive features with enough variation and fidelity. We will show our quantitative results based on classifiers and the utilization of the synthetic data set in experiments section in detail.

5.2. Text-Driven Medical Image Manipulated Augmentation

With the recent developments from OpenAI [56] , we decided to utilize the representational power of the CLIP pre-trained model in the context of latent space manipulation. Patashnik *et al.* [2] proposed a framework called StyleCLIP which enables semantic meaningful latent space manipulation with text guidance. In their framework, they worked on the latents spaces of StyleGAN, StyleGAN2 and StyleGAN2-ADA. They employed and experimented with three different approaches which are text-guided latent optimization, latent residual mapper, mapping a text prompt into an input agnostic global direction. In one hand, text-guided latent optimization works on W+ space and takes several minutes to find an optimal style manipulation direction vector for a given text and image pair. On the other hand, Latent residual mapper requires 10 to 12 hours for training the mapper network on a single NVIDIA GTX 1080Ti GPU. Latent residual mapper has the control over several attributes per text for an image. For instance, they were able to generate Angela Merkel's face image with curly and long hair by only inputting "Curly long hair" in the text field. In our approach, we did not experimented with both of these methods. Instead, we only experimented with the input agnostic global direction method which does not require further optimization. For a given pair of text which defines neutral and target attributes, global direction method exploits the colinearity between CLIP's image embedding space and StyleGAN's style space for the difference of the given text pair. In other words, they found that given a pair of text, the corresponding difference vector in the image embedding

space is colinear with the StyleGAN's style space. In our work, we used that property for finding disease specific difference vector to generate semantically meaningful manipulated images.

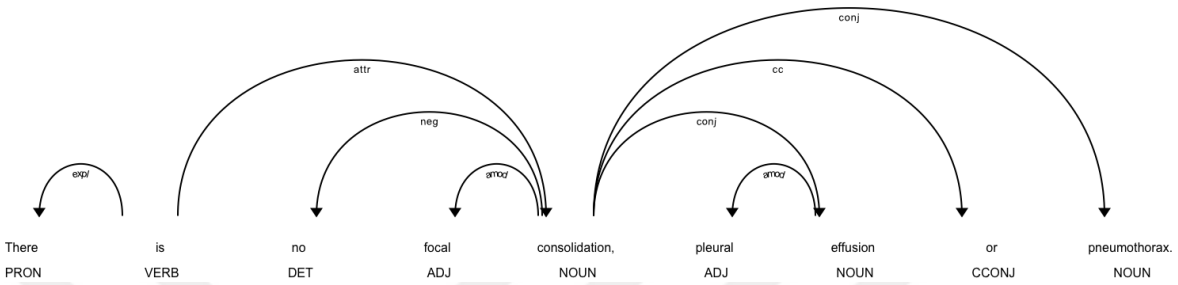

Figure 5.1. A sample dependency parser output of a medical report.

During our initial experiments on StyleCLIP with default CLIP weights, we saw that global optimization method does not yield desired image manipulations in the context of chest X-rays. We believed that this is due to lack of medical data in CLIP's training set. Thus, we fine-tuned CLIP with medical image - text pairs. Our simple yet effective fine-tunning approach yielded promising results. We improved fine-tuning process by introducing NLP techniques for medical reports. To create more informative reports, we extracted entities by using scispaCy [73] which contains spaCy [74] models for processing biomedical-specific textual data. In our experiments, we used the entities which is yielded by *en_core_sci_scibert* model. This model is a complete spaCy pipeline built on top of transformer and pre-trained on a large biomedical corpus consists of nearly 785 thousand words. In one hand we only used the entities extracted for each report instead of full report. On the other hand, we used dependency relations between words and we modified the entities to create less complex semantics for each report. For instance, for an X-ray image, the corresponding report is "There is no focal consolidation, pleural effusion or pneumothorax". The entities found by scispaCy model are "no focal consolidation", "pleural effusion" , "pneumothorax". The dependency relation graph in Figure 5.1 also yields us "No" is related to "no focal consolidation", "effusion" and "consolidation" directly. "No" is also related to "pneumothorax" indirectly. However, our rule-based approach yields us "no focal consolidation", "no pleural effusion" and "no pneumothorax" by combining the entities which are directly

or indirectly related to word “No”. If the entity includes “No”, we remove it from the entity. Of course “No” is not only the word that affects the semantic complexity of the medical reports. We analyzed the frequency of the words for all reports and formed a generic list of words that affect semantic complexity. In the end, we transformed all the entities for each report according to the generic list and fine-tuned the CLIP with modified entities-image pairs. We named this fine-tuning strategy as rule-based. However, the results of our experiments showed that rule-based approach performs poorly for some classes. To fix this issue, we conducted further analysis and discovered new indicator words. The fine-tuning strategy that includes this words in the generic list was named as rule-based-v2.

6. EXPERIMENTS

In this chapter, we covered the key evaluation metrics which we used in our research. Afterwards, we explained the experiments we conducted for both synthetic medical data augmentation and text-driven latent space manipulated GAN augmentation. For each experiment section, we reported, compared and discussed the results in detail.

6.1. Evaluation Metrics

In this section, key evaluation metrics are explained to convey better understanding of the results of our experiments. In short, Frechet Inception Distance (FID) is used for measuring the quality of synthetic images. ROC-AUC and PR-AUC scores are used for computing classifier's overall performance across all classes.

6.1.1. Frechet Inception Distance

It is a challenging task to measure the divergence and literalness of generated samples in the same time. However, to tackle this problem, Frechet Inception Distance (FID) was introduced by Heusel *et al.* [75] in order to quantitatively evaluate the synthetic images generated by generative models in terms of how realistic yet divergent they appear. FID metric was specifically developed for performance measurements of GANs. Unlike pixel-wise comparison between real and generated samples, roughly speaking, FID metric computes the distance between feature vectors of real images and feature vectors of generated images with the use of Inception V3 [76] architecture. Mostly it outputs closer results to human-grade evaluation compared to the prior evaluation metrics. FID was well-understood and adopted by most of the influencive approaches in the domain of GANs.

FID basically measures the distance of distributions for real and synthetic images. The distributions are obtained from the final pooling layer of Inception V3 model. FID is computed as follows:

$$\mathrm{FID}(r,g) = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\sum_r + \sum_g - 2\sqrt{\sum_r \sum_g}\right). \tag{6.1}$$

Mean and covariance of the real and generated data are $(\mu_r, \sum_r)$, $(\mu_g, \sum_g)$ respectively. Tr means trace linear algebra operation which is the summation of all diagonal entities of a given square matrix. Since FID computes the difference of the distributions for real and generated images, scores which are close to 0 means the generated samples are realistic. Therefore, during the performance comparison of the two different GANs approaches, we state the approach with lower FID score is better.

6.1.2. ROC-AUC

ROC is short for Receiver Operation Characteristic and AUC is the acronym for Area Under the Curve. ROC curve is used for building a better understanding of the classifier's performance. In order to plot ROC curve for a binary classifier, probability scores of a class are collected by feeding the classifier with a subset of the validation or the test data. After obtaining the classification scores, True Positive Rate and False Positive Rate is calculated for different decision thresholds. Each threshold corresponds to a True Positive Rate and False Positive Rate which are computed as follows:

$$TruePositiveRate = TruePositive/(TruePositive + FalseNegative), \tag{6.2}$$

$$FalsePositiveRate = FalsePositive/(FalsePositive + TrueNegative). \tag{6.3}$$

After sorting the list of pairs by False Positive Rates ascending, the projection of the data points yields us the ROC curve. Moreover, the calculation of the area under the ROC curve yields ROC-AUC score. ROC-AUC score is expected to be in between 0.0 and 1.0. If a classification result gives 1.0 ROC-AUC score for test data, that means the classifier works perfectly. However, if a classification result gives 0.0 ROC-AUC score, it means the classifier did not learn anything from the training data. See Figure

6.1 for a sample ROC curve. In this example, the AUC score is 0.80 which is a fairly good binary classification performance.

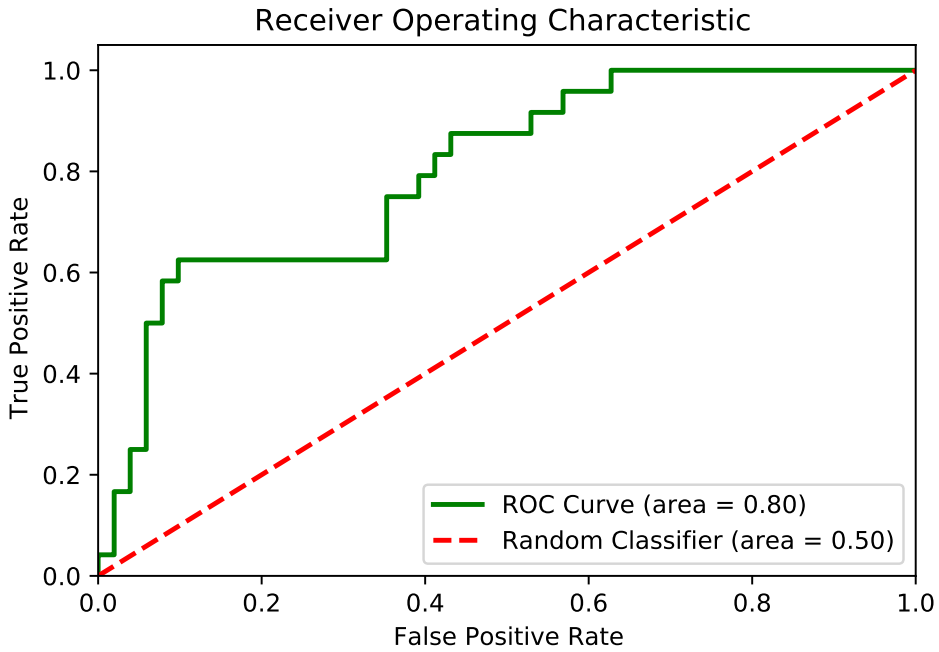

Figure 6.1. Receiver Operation Characteristic Curve graph for an example binary classifier.

Although ROC-AUC score indicates well the overall classification performance for balanced datasets, in case of inbalance between positive and negative samples, ROC-AUC score may perform poorly for the assessment of classification accuracy since it gives equal importance to positive classification performance and negative classification performance. If a dataset contains mostly negative samples, negative classification performance may dominate ROC-AUC score. The classifier trained with this kind of imbalanced dataset may have poor classification performance for positive classes.

6.1.3. PR-AUC

PR-AUC is the area under the precision-recall curve. PR curve is plotted nearly the same with ROC-AUC curve with a slight difference.

Instead of True Positive Rate and False Positive Rate, Precision and Recall metrics are used as axis. The equations of Precision and Recall are as follow:

$$Precision = TruePositive/(TruePositive + FalsePositive), \tag{6.4}$$

$$Recall = TruePositive/(TruePositive + FalseNegative). \tag{6.5}$$

Except with this minor change, PR curve is computed and plotted same as ROC curve. An example PR curve could be found in Figure 6.2. In this figure, an arbitrary classification model was used. AP value means average precision. Mathematically, it is also the same value as PR-AUC. AP and PR-AUC could be used interchangeably depending on the writers' convention. In Figure 6.2, we used AP instead of PR-AUC.

According to [77], PR-AUC metric is suitable for imbalanced datasets since it is more sensitive to the classification performance of minority classes compared to ROC-AUC metric. Therefore, we also report PR-AUC scores of our experiments.

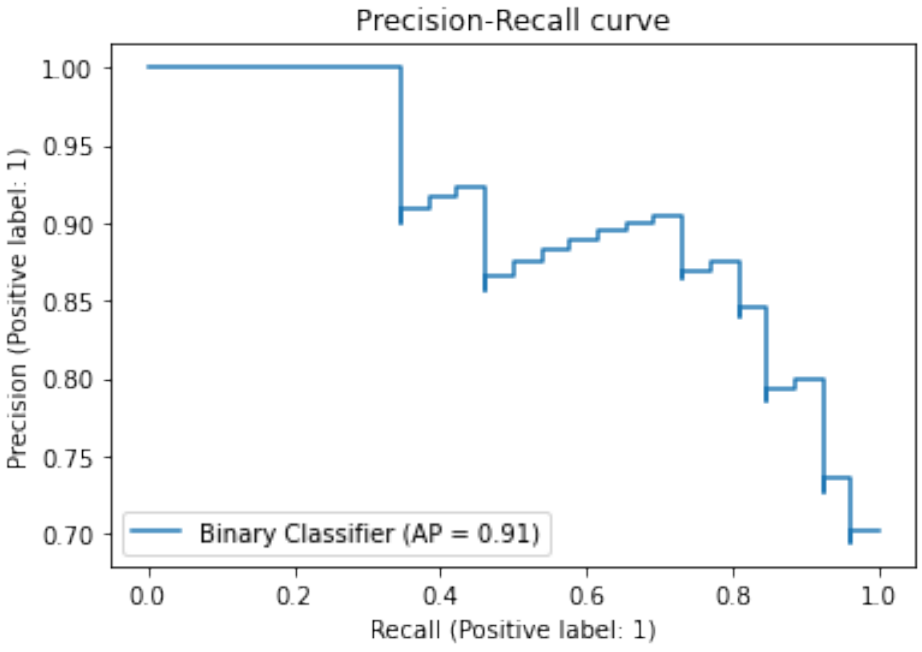

Figure 6.2. Precision Recall Curve graph for an example binary classifier.

6.2. Synthetic Medical Data Augmentation

CheXpert dataset is a relatively large dataset compared to other medical radiographic datasets. We used the down-sampled version of CheXpert dataset which consists of nearly 225.000 chest radiographs of the patients who underwent medical examinations from Stanford University Medical Center. The dataset includes samples of 14 different classes with uncertain labels. Each sample in the dataset is labelled either positive(1), negative(0) or uncertain(-1). In our experiments we focused on 5 classes which are Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion since the classification in between these classes is the subject of the CheXpert competition. We removed the samples with multi-class-labels from the dataset and used the samples with only positive or negative labels. That yielded us a dataset with nearly 85.000 samples. Since StyleGAN2-ADA only performs on square images with the size of powers of 2, we resized the samples to 256×256 resolution with black padding. We trained StyleGAN2-ADA with default configurations in class conditional setting on 4 NVIDIA Tesla V100 GPUs. We saved the model for each 100 training iteration. After two days of training, we selected the best performing model which has nearly 17 FID score. We generated 5000 samples for each class. In Figure 6.3 an example class conditional generated sample for cardiomegaly class could be found.

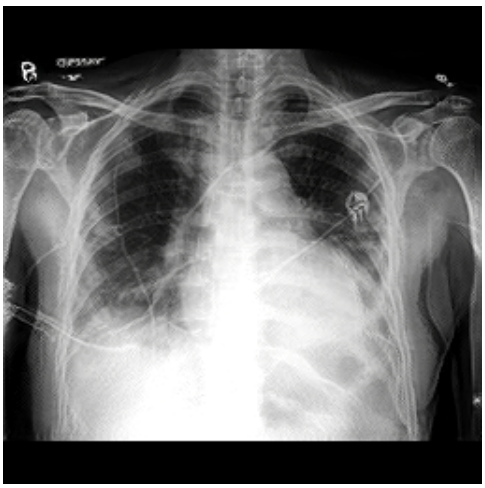

Figure 6.3. The class conditioned generated sample for cardiomegaly.

Table 6.1. The results of classification experiments with PCAM.

Augmentation	**Cardiomegaly (AUC)**	**Edema (AUC)**	**Consolidation (AUC)**	**Atelectasis (AUC)**	**Pleural Effusion (AUC)**	**Mean AUC**
No Aug (Original)	0,789	0,903	0,887	**0,872**	0,883	0,867
5k	0,820	**0,904**	0,890	0,853	0,912	0,876
10k	0,838	0,903	0,927	0,854	**0,915**	**0,887**
15k	**0,845**	0,895	0,926	0,847	0,887	0,880
20k	0,842	0,894	**0,928**	0,828	0,911	0,881

In our experiments, we trained PCAM framework with original dataset in order to provide a strong baseline for our augmentation pipeline. We also extended the original dataset with different numbers of additional generated samples. To observe the correlation between the number of additional synthetic samples and the improvement on classification accuracy, we added 5000, 10000, 15000, 20000 samples per each class. We reported the results at Table 6.1 for our initial experiments. Results showed that the amount of augmentation affects classification performance depending on the class. For instance, The augmentation of 20k additional samples yielded best result for Consolidation. However, additional data injection degraded the performance of Atelectasis classification. If we look at the big picture, 10K additional samples per each class performs best for overall classification.

In our following experiments, we are aiming to augment the dataset with class specific numbers of generated samples to solve the class imbalance problem. In CheXpert dataset analysis, we saw that the minority class has nearly 4000 samples whereas the majority class has nearly 35000 samples. Therefore we aimed to balance the dataset with certain number of generated classes. We also changed the classifier due to the classical augmentation techniques they used in their implementation. Moreover, we run the new classifier on our generated samples to understand our class conditional generation performance. After understanding the feature representational power of the generated dataset, we balanced the dataset with certain levels of generated images and we compared the results with their trained model without GAN augmentation. The classification results for only synthetic dataset with 5000 samples per medical con-

dition, semi-balanced dataset with 16000 samples for each minority classes and bare dataset (original dataset without any additional synthetic sample) could found in Table 6.2.

Table 6.2. GAN augmentation experiment results in terms of ROC-AUC.

Medical Condition	**Synthetic Only**	**Bare Dataset**	**Semi-balanced Dataset**
Atelectasis	0,490	**0,847**	0,839
Cardiomegaly	0,667	**0,831**	0,821
Consolidation	0,592	0,899	**0,933**
Edema	0,715	0,871	**0,897**
Pleural Effusion	0,649	0,900	**0,925**
AUROC mean	0,623	0,870	**0,883**

6.3. Text-Driven Latent Space Manipulation

In this section, we used StyleCLIP framework with global direction method. To conduct experiments with this approach, we needed two different pre-trained models which are StyleGAN2-ADA with conditional setting (cifar10c) and CLIP model. In our initial experiments, we used the original CLIP model trained with almost 400 million text-image pair. However, we could not obtain any meaningful style transfer output since the CLIP model is not well suited for chest X-ray image domain. In order to adopt CLIP's domain to chest X-ray image domain, we fine-tuned pre-trained CLIP ViT-B/32 model with MIMIC-CXR dataset since the reports of CheXpert dataset is not currently available. Because of the fact that the original CLIP model restraints maximum positional embedding size with 77, we trimmed the medical report texts accordingly. After fine-tuning, we repeated the experiment. Our results could be found at Figure 6.4 and 6.5.

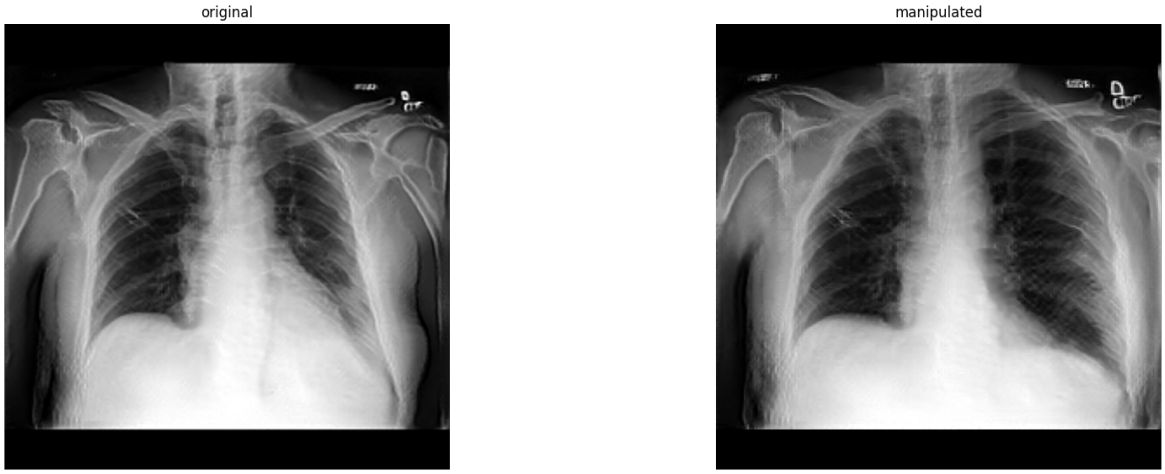

Figure 6.4. Experiment for neutral text normal and target text lung volumes.

To quantify the augmentation effects of text-driven style manipulation, we sampled 100 images for each particular medical condition (cardiomegaly, pleural effusion, edema, consolidation and atelectasis) from CheXpert dataset. We inverted the samples using e4e to get the latent variables which generates almost identical synthetic image of the original image. We manipulated each inverted image depending its medical condition by using global optimisation method of StyleCLIP. For instance, we applied no cardiomegaly to cardiomegaly manipulation direction to the inverted samples which are originated from cardiomegaly tagged X-ray samples.

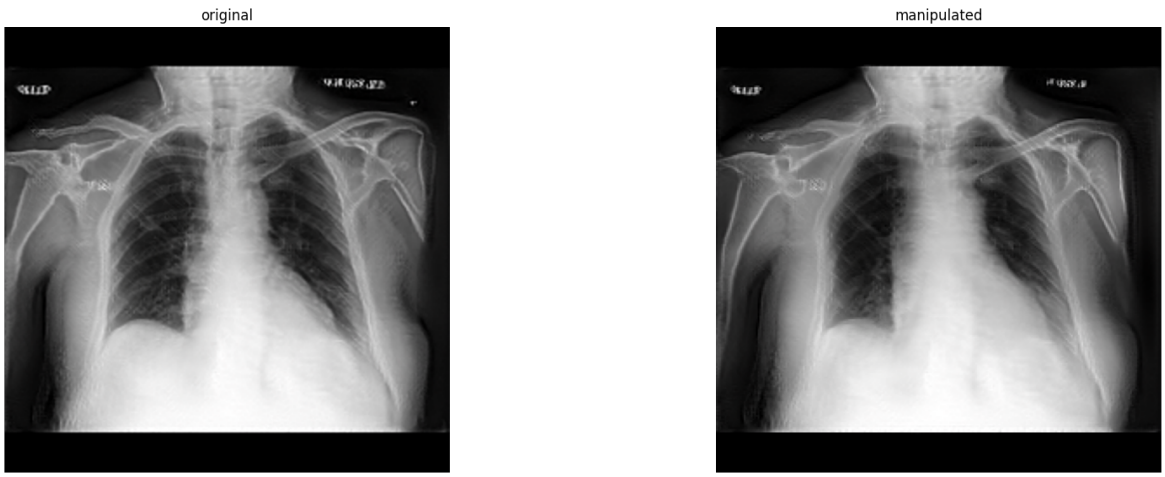

Figure 6.5. Experiment for neutral text normal and target text cardiomegaly.

To obtain meaningful manipulations, it is important to adjust disentanglement threshold and manipulation strength. Disentanglement threshold determines if the manipulation alters course features or fine-grained features. Manipulation straight controls the degree of the manipulation. In our experiment, we adjusted disentanglement threshold for each medical condition to ensure the manipulation does not distort the overall structure of the each inverted X-ray sample. In other words, we set the manipulation threshold empirically. After finding the optimal manipulation threshold value, we applied negative and positive values of manipulation strength to all the samples. While positive values increase the target feature, negative values decreases the target feature on an inverted image. To be more concrete, while we were manipulating inverted samples of the cardiomegaly class from no cardiomegaly to cardiomegaly direction, we observed that using positive values of manipulation strength enlarges the heart on the synthetic X-ray image. When we used negative values we saw that the heart on the synthetic X-ray image was shrunken.

To asses the effects of the manipulations in our preliminary experiments, we used a classifier which is trained on CheXpert dataset. We labelled each image according to its original label and run the classification for four different dataset. The first experiment consists of the original images only. The second one contains only inverted samples. The third one includes manipulated images only with positive manipulation strength. The final experiment consists of manipulated images only with negative manipulation strength.

To understand the original CLIP model's attribute based manipulation capabilities, we conducted an experiment with untouched CLIP model. In this experiment we used 100 samples for each class. Results could be seen in Table 6.3. The results showed that style manipulations with original CLIP model does not control disease attributes well except for two diseases which are cardiomegaly and consolidation.

Table 6.3. The classification results of style manipulations for original CLIP model for 100 samples.

	Original X-Rays (AUC)	Negative Directions X-Rays (AUC)	Inverted X-Rays (AUC)	Positive Directions X-Rays (AUC)
Atelectasis	**0,6054**	0,5647	0,5116	0,5440
Cardiomegaly	0,7317	0,3846	0,5749	**0,8353**
Consolidation	**0,6774**	0,4513	0,4687	0,5143
Edema	**0,8448**	0,7802	0,6459	0,5756
Pleural_Effusion	**0,8208**	0,4659	0,4931	0,4719
AUC mean	**0,7360**	0,5293	0,5388	0,5882

The results are shown in Table 6.4 is for fine-tuned CLIP for 100 samples. Example images could also could be found in Figure 6.6. According to the results, our approach managed to control disease specific features in most of the cases except consolidation and edema classes. We believed that lack of inversion performance affects these classes. On the other hand, we confirmed that we are able to control the disease specific features by considering AUC means of all experiments since the AUC mean of positive directions is greater than the AUC mean of only inverted images. AUC mean of only inverted images are also greater than negative directions which confirms that we are also able to extract the disease specific features from the inverted sample. In our next experiment, we manipulated the attributes for 1000 samples to develop better understanding of the style manipulations. In Table 6.5 the results could be found. According to the results, attribute manipulations for atelectasis, cardiomegaly and pleural effusion is well captured by our approach.

Table 6.4. The classification results of style manipulations for fine-tuned CLIP for 100 samples.

	Original X-Rays (AUC)	**Negative Directions X-Rays (AUC)**	**Inverted X-Rays (AUC)**	**Positive Directions X-Rays (AUC)**
Atelectasis	**0,6054**	0,4287	0,5116	0,5620
Cardiomegaly	0,7317	0,5255	0,5749	**0,7652**
Consolidation	**0,6774**	0,4893	0,4687	0,3803
Edema	**0,8448**	0,5499	0,6459	0,5390
Pleural_Effusion	0,8208	0,4860	0,4931	**0,8472**
AUC mean	**0,7360**	0,4959	0,5388	0,6187

Table 6.5. The classification results of style manipulations for fine-tuned CLIP for 1000 samples.

	Original X-Rays (AUC)	Negative Directions X-Rays (AUC)	Inverted X-Rays (AUC)	Positive Directions X-Rays (AUC)
Atelectasis	0,5835	0,3771	0,5101	**0,6085**
Cardiomegaly	0,7545	0,4026	0,5964	**0,7858**
Consolidation	**0,6585**	0,4972	0,5174	0,4502
Edema	**0,7907**	0,5446	0,6551	0,3745
Pleural_Effusion	0,8168	0,5857	0,5723	**0,8340**
AUC mean	**0,7208**	0,4814	0,5703	0,6106

For our first style based manipulation augmentation experiment, we trained the classifier for the bare CheXpert dataset, GAN augmented CheXpert dataset and style manipulation based GAN augmented Chexpert dataset. The augmentations for two augmented datasets which are GAN (StyleGAN2-ADA class conditional generation) augmented and style based GAN augmented datasets includes 1k extra samples for each class. Bare dataset accuracies are shown as baseline results since bare dataset does not contain any kinds of augmentations. The results from Table 6.6 show us style manipulation augmentation method performs better for Atelectasis, Cardiomegaly and Edema classes. Moreover, considering overall performance, style manipulation based GAN augmentation technique outperforms conventional GAN augmentation method.

Table 6.6. Classifier training accuracies with respect to bare dataset, GAN augmented and style manipulation based GAN augmentation.

	Bare Dataset (AUC)	GAN Augmentation (AUC)	Style Manipulation Augmentation (AUC)
Atelectasis	0,84716	0,81834	**0,82760**
Cardiomegaly	0,83088	0,79979	**0,81556**
Consolidation	0,89929	**0,90532**	0,89809
Edema	0,87125	0,88912	**0,90218**
Pleural_Effusion	0,89981	**0,92671**	0,91474
AUC mean	0,86968	0,86786	**0,87163**

After obtaining the preliminary results which are promising, we expended our experiments with various CLIP fine-tuning strategies. We used impressions section of MIMIC-CXR dataset for fine-tuning. We also used WGSum framework [78] for impression generation and used the generated impressions for the fine-tuning. Moreover, we analyzed our rule-based approach in order the improve the entanglement of the disease specific features. Especially, our rule-based approach is not able to generate semantically meaningful style manipulation directions for Edema and Consolidation classes. To resolve this issue, we improved our rule-based approach and named it Rule-Based-v2. We generated style manipulated images for both positive and negative directions to see if the manipulation direction is able to control the disease specific attributes. In order to assess style manipulations with a classifier, we utilized DeepAUC framework [79] which is the top solution of the CheXpert competetion. We trained the classifier with bare dataset without any built-in augmentations or data enhancement strategies to see the pure effect of our approach. Finally, we trained the classifier with augmented datasets to evaluate the results. Since the CheXpert dataset is imbalanced, we reported PR-AUC scores along with ROC-AUC scores to examine the effects of augmentations independently. This allows us to compare the difference between ROC-AUC scores and PR-AUC scores.

The fine-tuning strategies we applied in our experiments are:

- rule-based: The CLIP we used in StyleCLIP for sample generation is fine-tuned with the outputs of our rule-based POS tagging algorithm and corresponding X-ray images. Findings sections of MIMIC-CXR reports are used as input.
- impression: The CLIP we used in StyleCLIP for sample generation is fine-tuned with impressions sections of MIMIC-CXR radiology reports and corresponding X-ray images.
- rule-based-V2: The CLIP we used in StyleCLIP for sample generation is fine-tuned with the outputs of our newer version of rule-based POS tagging algorithm and corresponding X-ray images. Findings sections of MIMIC-CXR reports are used.

- original-CLIP: The CLIP we used in StyleCLIP for sample generation is untouched.
- randomized: The CLIP we used in StyleCLIP for sample generation is fine-tuned with MIMIC-CXR reports with random word order and corresponding X-ray images. Findings sections of MIMIC-CXR reports are used as input.
- inverted: The CLIP we used in StyleCLIP for sample generation is fine-tuned with inverted X-ray images and corresponding radiology reports. Findings sections of MIMIC-CXR reports are used as input.
- WGSum-generated: The CLIP we used in StyleCLIP for sample generation is fine-tuned with the WGSum-generated summaries of radiology reports and corresponding images. Findings sections of MIMIC-CXR reports are used as input.

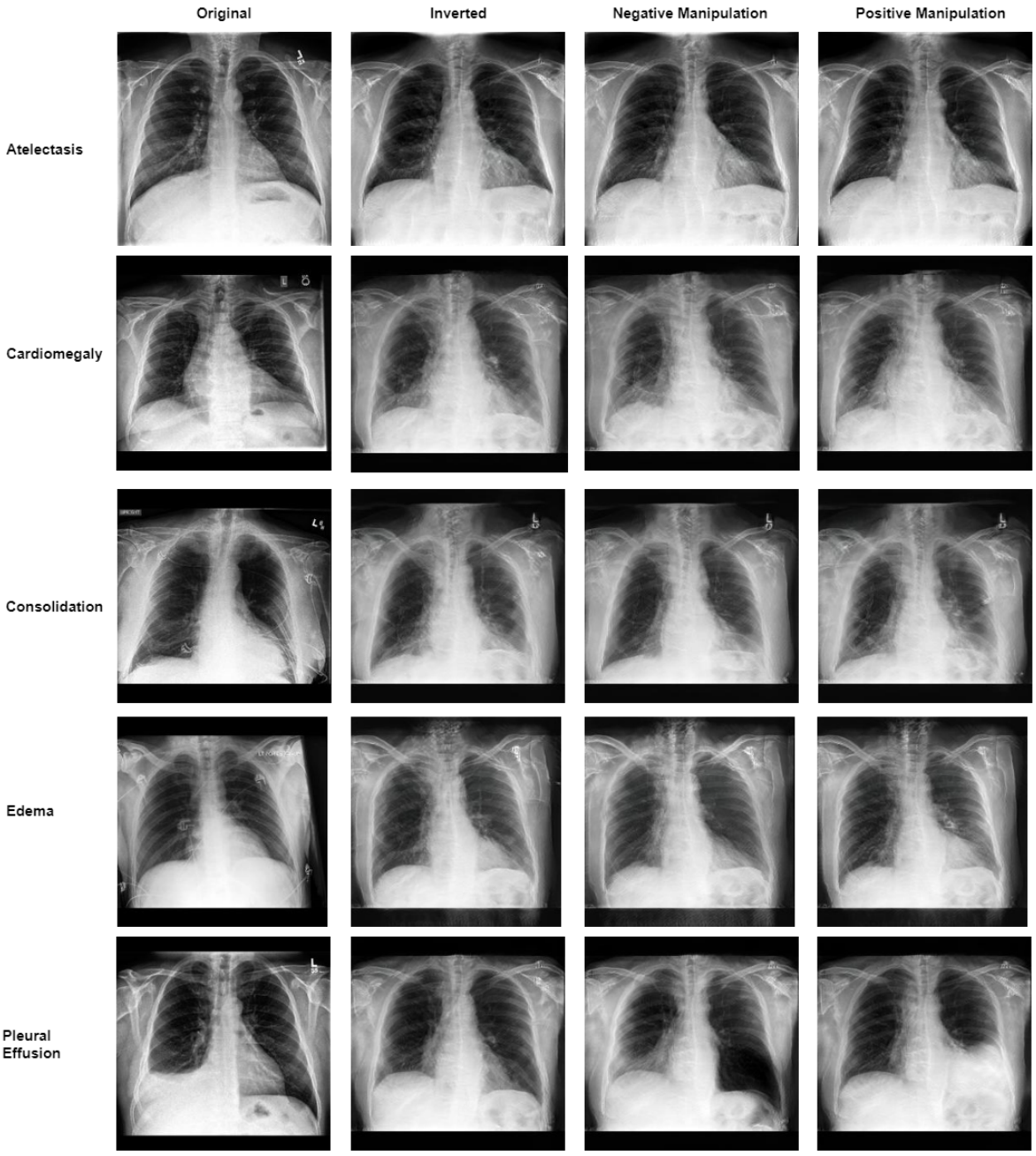

Figure 6.6. Original, inverted and manipulated images with positive and negative directions generated by StyleCLIP with fine-tuned CLIP model.

Table 6.7. Classification results on only synthetic style manipulated data in terms of ROC-AUC score.

Fine-Tuning Strategy	**Cardiomegaly (ROC-AUC)**	**Edema (ROC-AUC)**	**Consolidation (ROC-AUC)**	**Atelectasis (ROC-AUC)**	**Pleural Effusion (ROC-AUC)**	**Mean AUC**
rule-based (-)	0,366	0,526	0,471	0,493	0,450	0,4611
impression (-)	0,459	0,561	0,411	0,448	0,438	0,4633
rule-based-V2 (-)	0,525	0,605	0,233	0,460	0,565	0,4776
original-CLIP (-)	0,385	0,717	0,518	0,493	0,668	0,5562
WGSum-generated (-)	0,650	0,602	0,548	0,487	0,506	0,5586
randomized (-)	0,442	0,634	0,445	0,559	0,664	0,5485
inverted	*0,535*	*0,637*	*0,542*	*0,504*	*0,571*	*0,5576*
rule-based (+)	**0,709**	0,605	0,451	0,626	0,783	0,6347
impression (+)	0,600	0,566	**0,748**	**0,639**	0,772	0,6649
rule-based-V2 (+)	0,507	0,799	0,583	0,635	**0,803**	**0,6654**
original-CLIP (+)	0,656	0,731	0,724	0,559	0,518	0,6377
WGSum-generated (+)	0,328	**0,846**	0,635	0,567	0,636	0,6025
randomized (+)	0,637	0,487	0,661	0,524	0,477	0,5570

In the Table 6.7, the dataset contains 1k generated X-ray image per each class. The dataset includes 5k synthetic data in total. Minus sign denotes negative manipulation direction. Plus sign means positive manipulation direction. According to Table 6.7, positive direction of style manipulated synthetic images generated by rule-based method impose the best attribute control for Cardiomegaly specific attributes. WGSum-generated images with the positive style manipulation direction perform better than any other method for Edema. For Consolidation and Atelectasis, positive directional style manipulations outperform other style manipulation strategies. In case of Pleural Effusion, rule-based-V2 with positive direction is the top performing style manipulation strategy for the enhancement of the disease specific attributes. If we look at overall attribute enhancement performance, rule-based-V2 is the top performing manipulation strategy in terms of ROC-AUC score.

Table 6.8 shows the classification results in terms of PR-AUC score. Results are mostly parallel to Table 6.7 considering the attribute enhancement ability. However, the PR-AUC scores are rather lower than the ROC-AUC scores since the training and the validation set of CheXpert dataset is imbalanced. Another significant difference is the mean AUC score. According to PR-AUC scores, impression with positive direction

performs best at attribute enhancement for all classes combined. However, according to the ROC-AUC scores, the overall attribute enhancement ability of rule-based-v2 technique outperforms other techniques. If we have to choose a single manipulation strategy for enhancing disease specific features , we should go for impressions with positive direction method instead of rule-based-V2.

For tables 6.9 to 6.12, we present the results of style manipulated GAN augmentations. Since the dataset is imbalanced, during the training when the classification performance for a fine-tuning strategy reach the maximum mean ROC-AUC score, it does not need to reach the maximum mean PR-AUC score. This statement is also valid for vice versa. Therefore, we present two pairs of result tables for two different states of the classifier. One state is when the mean ROC-AUC score is maximum for each fine-tuning strategy. The other state is when the mean PR-AUC score is maximum for each fine-tuning strategy.

Our style manipulated GAN augmentation experiments cover both single training strategy and multiple fine-tuning strategies combined. The last three lines on the result tables show the combination of multiple fine-tuning strategies. Specifically, ensemble covers the maximum PR-AUC scored generated samples from various training strategies. To select the data we used Table 6.8. For instance, we included synthetic Cardiomegaly samples originated from rule-based(+) method. We injected synthetic Edema samples from WGSum-Generated(+) method. Since there may not be available impressions section for any other dataset, we employed ensemble-without-impression method. In this method, we ignored the data created by impression fine-tuning strategy and select the data with maximum PR-AUC score. For instance, we included data created by original-CLIP (+) method for Consolidation. We also selected data generated by rule-based-V2(+) method instead of impression (+) method. Considering Table 6.12, our data gathering strategy performs the best for overall classification in terms of PR-AUC score.

Table 6.8. Classification results on only synthetic style manipulated data in terms of PR-AUC score.

Fine-Tuning Strategy	**Cardiomegaly (PR-AUC)**	**Edema (PR-AUC)**	**Consolidation (PR-AUC)**	**Atelectasis (PR-AUC)**	**Pleural Effusion (PR-AUC)**	**Mean AUC**
rule-based (-)	0,147	0,198	0,196	0,194	0,193	0,186
impression (-)	0,191	0,203	0,159	0,169	0,175	0,179
rule-based-V2 (-)	0,234	0,228	0,126	0,169	0,227	0,197
original-CLIP (-)	0,151	0,410	0,225	0,197	0,411	0,279
WGSum-generated (-)	0,254	0,241	0,207	0,247	0,240	0,238
randomized (-)	0,190	0,269	0,182	0,261	0,312	0,243
inverted	*0,241*	*0,294*	*0,219*	*0,221*	*0,277*	*0,251*
rule-based (+)	**0,579**	0,235	0,170	0,364	0,432	0,356
impression (+)	0,270	0,232	**0,353**	**0,516**	0,455	**0,365**
rule-based-V2 (+)	0,208	0,382	0,222	0,281	**0,619**	0,342
original-CLIP (+)	0,339	0,380	0,336	0,260	0,205	0,304
WGSum-generated (+)	0,159	**0,579**	0,281	0,221	0,315	0,311
randomized (+)	0,318	0,181	0,288	0,217	0,187	0,238

Table 6.9. Classification results of fine-tuning strategies on augmented dataset in terms of ROC-AUC score. Each line was obtained when the mean ROC-AUC score is maximum.

Fine-Tuning Strategy	**Cardiomegaly (ROC-AUC)**	**Edema (ROC-AUC)**	**Consolidation (ROC-AUC)**	**Atelectasis (ROC-AUC)**	**Pleural-Effusion (ROC-AUC)**	**Mean AUC**
pure-dataset	0,836	0,928	0,919	0,843	0,924	0,890
impression (+)	0,826	0,913	**0,936**	0,844	0,923	0,888
WGSum-generated (+)	0,846	0,918	0,911	0,852	0,927	0,891
rule-based (-)	0,860	0,923	0,910	0,842	0,927	0,892
impression (-)	0,868	0,920	0,913	0,847	0,933	0,896
original-clip (-)	**0,871**	0,928	0,920	0,828	0,932	0,896
original-clip (+)	0,845	0,919	0,935	0,860	0,918	0,896
rule-based (+)	0,867	0,928	0,926	0,828	0,929	0,896
rule-based (+) and inverted	0,840	**0,938**	0,929	0,846	0,925	0,896
randomized (+)	0,859	0,927	0,924	0,850	0,926	0,897
WGSum-generated (-)	0,866	0,918	0,930	0,834	0,937	0,897
rule-based (+), inverted and rule-based-V2(+)	0,835	0,924	0,923	**0,877**	0,927	0,897
ensemble	0,866	0,926	0,905	0,846	**0,940**	0,897
inverted	0,867	0,927	0,928	0,835	0,937	0,899
randomized (-)	0,859	0,930	0,932	0,846	0,935	0,900
rule-based-V2 (+)	0,866	0,924	0,931	0,844	0,935	0,900
ensemble-without-impression	0,863	0,927	0,929	0,854	0,929	0,900
rule-based-V2 (-)	0,857	0,933	0,929	0,847	0,937	**0,901**

Table 6.10. Classification results of fine-tuning strategies on augmented dataset in terms of PR-AUC score. Each line was obtained when the mean ROC-AUC score is maximum.

Fine-Tuning Strategy	**Cardiomegaly (PR-AUC)**	**Edema (PR-AUC)**	**Consolidation (PR-AUC)**	**Atelectasis (PR-AUC)**	**Pleural-Effusion (PR-AUC)**	**Mean AUC**
pure-dataset	0,716	0,790	0,639	0,696	0,850	0,738
WGSum-generated (+)	0,680	0,751	0,583	0,750	0,853	0,723
impression (+)	0,717	0,743	0,642	0,674	0,858	0,727
ensemble	0,763	0,777	0,526	0,696	0,873	0,727
inverted	0,756	0,772	**0,568**	0,682	0,873	0,730
rule-based (+)	0,747	0,784	0,587	0,682	0,856	0,731
rule-based (-)	0,746	0,767	0,582	0,704	0,859	0,732
impression (-)	**0,772**	0,747	0,553	0,736	0,867	0,735
ensemble-without-impression	0,747	0,773	0,569	0,730	0,865	0,737
original-clip (-)	0,763	0,782	0,580	0,694	0,868	0,737
original-clip (+)	0,749	0,760	0,648	0,684	0,846	0,737
rule-based-V2 (-)	0,740	0,777	0,588	0,703	**0,876**	0,737
WGSum-generated (-)	0,762	0,755	0,608	0,689	0,871	0,737
randomized (-)	0,749	0,784	0,578	0,723	0,873	0,742
randomized (+)	0,733	0,784	0,610	0,726	0,861	0,743
rule-based-V2 (+)	0,760	0,773	0,593	0,730	0,872	0,746
rule-based (+), inverted and rule-based-V2(+)	0,738	0,788	0,614	**0,786**	0,832	0,752
rule-based (+) and inverted	0,723	**0,796**	0,658	0,743	0,847	**0,753**

Table 6.11. Classification results of fine-tuning strategies on style manipulated GAN augmented dataset in terms of ROC-AUC score. Each line was obtained when the mean PR-AUC score is maximum.

Fine-Tuning Strategy	**Cardiomegaly (ROC-AUC)**	**Edema (ROC-AUC)**	**Consolidation (ROC-AUC)**	**Atelectasis (ROC-AUC)**	**Pleural-Effusion (ROC-AUC)**	**Mean AUC**
pure-dataset	0,831	0,930	0,920	0,802	0,912	0,879
rule-based (-)	0,755	0,930	0,932	0,818	0,910	0,869
WGSum-generated (-)	0,770	0,927	0,920	0,825	0,907	0,870
rule-based-V2 (+)	0,779	0,930	0,924	0,821	0,908	0,873
impression (-)	0,775	0,931	0,939	0,829	0,893	0,873
rule-based-V2 (-)	0,770	0,929	0,932	0,837	0,907	0,875
rule-based (+)	0,793	0,929	0,934	0,830	0,902	0,877
inverted	0,813	0,921	0,926	0,818	0,913	0,878
ensemble-without-impression	0,802	0,926	0,930	0,838	0,897	0,879
randomized (+)	0,798	0,934	0,909	0,851	0,905	0,879
randomized (-)	0,795	0,923	**0,941**	0,839	0,901	0,880
ensemble	0,780	0,931	0,939	0,836	0,916	0,880
original-clip (-)	0,781	0,933	0,939	**0,852**	0,903	0,882
rule-based (+), inverted and rule-based-V2(+)	0,833	0,922	0,930	0,795	**0,929**	0,882
impression (+)	0,829	0,933	0,894	0,827	0,927	0,882
original-clip (+)	**0,843**	0,933	0,894	0,824	0,923	0,884
WGSum-generated (+)	0,831	0,934	0,921	0,830	0,925	0,888
rule-based (+) and inverted	0,840	**0,938**	0,929	0,846	0,925	**0,896**

Table 6.12. Classification results of fine-tuning strategies on style manipulated GAN augmented dataset in terms of PR-AUC score. Each line was obtained when the mean PR-AUC score is maximum.

Fine-Tuning Strategy	**Cardiomegaly (PR-AUC)**	**Edema (PR-AUC)**	**Consolidation (PR-AUC)**	**Atelectasis (PR-AUC)**	**Pleural-Effusion (PR-AUC)**	**Mean AUC**
pure-dataset	0,719	0,745	0,615	0,633	0,842	0,711
rule-based (+), inverted and rule-based-V2(+)	0,707	0,756	0,592	0,644	0,847	0,709
impression (+)	0,688	0,766	0,562	0,680	**0,855**	0,710
original-clip (+)	**0,738**	0,779	0,566	0,667	0,843	0,719
rule-based-V2 (+)	0,658	0,800	0,643	0,660	0,841	0,720
rule-based (-)	0,641	0,782	0,728	0,624	0,837	0,722
inverted	0,700	0,788	0,660	0,651	0,841	0,728
rule-based (+)	0,684	0,796	0,659	0,693	0,821	0,731
WGSum-generated (-)	0,659	0,791	0,671	0,698	0,834	0,731
randomized (+)	0,678	**0,818**	0,614	0,712	0,835	0,731
rule-based-V2 (-)	0,649	0,787	0,703	0,707	0,838	0,737
randomized (-)	0,682	0,791	0,688	0,701	0,827	0,738
WGSum-generated (+)	0,687	0,790	0,669	0,691	0,852	0,738
ensemble-without-impression	0,694	0,796	0,690	0,701	0,818	0,740
impression (-)	0,664	0,796	0,758	0,682	0,823	0,744
original-clip (-)	0,670	0,810	0,706	0,727	0,827	0,748
rule-based (+) and inverted	0,723	0,796	0,658	**0,743**	0,847	0,753
ensemble	0,672	0,802	**0,783**	0,667	0,845	**0,754**

For tables 6.9 to 6.12, we present the results of style manipulated GAN augmentations. Augmented dataset covers full CheXpert dataset and the corresponding generated data depending on the fine-tuning strategy. Pure-dataset is present for benchmarking. Minus sign denotes negative manipulation direction. Plus sign means positive manipulation direction. Since the dataset is imbalanced, during the training when the classification performance for a fine-tuning strategy reach the maximum mean ROC-AUC score, it does not need to reach the maximum mean PR-AUC score. This statement is also valid for vice versa. Therefore, we present two pairs of result tables for two different states of the classifier. One state is when the mean ROC-AUC score is maximum for each fine-tuning strategy. The other state is when the mean PR-AUC score is maximum for each fine-tuning strategy.

Our style manipulated GAN augmentation experiments cover both single fine-tuning strategy and multiple fine-tuning strategies combined. Lines with multiple fine-tuning strategy names on the result tables show the combination of multiple fine-tuning strategies. To be more concrete, "rule-based (+) and inverted" strategy covers the generated data which are generated by both rule-based (+) and inverted fine-tunning strategy. However, ensemble strategies mix generated data without combining all the data. Specifically, "ensemble" covers the maximum PR-AUC scored generated samples from various training strategies. To select the data we used Table 6.8. For instance, we included synthetic Cardiomegaly samples originated from rule-based(+) method. We injected synthetic Edema samples from WGSum-Generated(+) method. Since there may not be available impressions section for any other dataset, we employed ensemble-without-impression method. In this method, we ignored the data created by impression fine-tuning strategy and select the data with maximum PR-AUC score. For instance, we included data created by original-CLIP (+) method for Consolidation. We also selected data generated by rule-based-V2(+) method instead of impression (+) method. Considering Table 6.12, our data gathering strategy performs the best for overall classification in terms of PR-AUC score.

7. CONCLUSION

Conventional data augmentation strategies broadly utilized by many image classification applications without sufficient training samples. These data augmentation methods comprises of yet not limited with reflection, arbitrary cropping, re-scaling existing samples and changes. These procedures are utilized during training the classifiers with augmented forms of real-world datasets. Expanding dataset with synthetically generated data samples permits us to improve the overall accuracy, decline overfitting and settle the training. With the substantial representation abilities of GANs, learning the distribution of real-world data with a reliable degree of variance allows us to produce image data with almost-in-secret discriminative features. In our methodology we utilized the previously mentioned generative ability of GANs by using the state-of-the-art GANs framework named as StyleGAN2-ADA. After the training of SytleGAN2-ADA in class conditional setting, we expanded the dataset size with various quantities of extra synthetic samples to explore the connection of classification performance and augmentation strength. We also introduced text-based style manipulated GAN augmentation technique for the medical domain. We utilized DeepAUC, the top solution of CheXpert competetion, to show the effectiveness of our GAN augmentation techniques. In our methodology, we saw that the classification performance of text-based manipulated GAN augmentation is better than classical GAN augmentation techniques.

REFERENCES

1. Karras, T., M. Aittala, J. Hellsten, S. Laine, J. Lehtinen and T. Aila, "Training Generative Adversarial Networks with Limited Data", *arXiv Preprint arXiv:2006.06676*, 2020.

2. Patashnik, O., Z. Wu, E. Shechtman, D. Cohen-Or and D. Lischinski, "Styleclip: Text-Driven Manipulation of Stylegan Imagery", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021.

3. Zhou, B., A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

4. Sampath, V., I. Maurtua, J. J. A. Martín and A. Gutierrez, "A Survey on Generative Adversarial Networks for Imbalance Problems in Computer Vision Tasks", *Journal of Big Data*, Vol. 8, No. 1, pp. 1–59, 2021.

5. Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren and A. Y. Ng, "Chexpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 590–597, 2019.

6. Hinton, G. E., "Boltzmann Machine", *Scholarpedia*, Vol. 2, No. 5, p. 1668, 2007.

7. Kingma, D. P. and M. Welling, "Auto-Encoding Variational Bayes", *arXiv Preprint arXiv:1312.6114*, 2013.

8. Pearl, J., "Bayesian Netwcrks: A Model Cf Self-Activated Memory for Evidential Reasoning", *Proceedings of the 7th Conference of the Cognitive Science Society,*

University of California, Irvine, CA, USA, pp. 15–17, 1985.

9. Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets", *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672–2680, 2014.

10. Alqahtani, H., M. Kavakli-Thorne and G. Kumar, "Applications of Generative Adversarial Networks (Gans): An Updated Review", *Archives of Computational Methods in Engineering*, pp. 1–28, 2019.

11. Mirza, M. and S. Osindero, "Conditional Generative Adversarial Nets", *arXiv Preprint arXiv:1411.1784*, 2014.

12. Radford, A., L. Metz and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", *arXiv Preprint arXiv:1511.06434*, 2015.

13. Denton, E., S. Chintala, A. Szlam and R. Fergus, "Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks", *arXiv Preprint arXiv:1506.05751*, 2015.

14. Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, "Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets", *arXiv Preprint arXiv:1606.03657*, 2016.

15. Zhao, J., M. Mathieu and Y. LeCun, "Energy-Based Generative Adversarial Network", *arXiv Preprint arXiv:1609.03126*, 2016.

16. Arjovsky, M., S. Chintala and L. Bottou, "Wasserstein Generative Adversarial Networks", D. Precup and Y. W. Teh (Editors), *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 214–223, PMLR, 2017.

17. Berthelot, D., T. Schumm and L. Metz, "Began: Boundary Equilibrium Generative Adversarial Networks", *arXiv Preprint arXiv:1703.10717*, 2017.

18. Karras, T., T. Aila, S. Laine and J. Lehtinen, "Progressive Growing of Gans for Improved Quality, Stability, and Variation", *arXiv Preprint arXiv:1710.10196*, 2017.

19. Brock, A., J. Donahue and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", *arXiv Preprint arXiv:1809.11096*, 2018.

20. Karras, T., S. Laine and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

21. Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and Improving the Image Quality of Stylegan", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

22. Bowles, C., L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw and D. Rueckert, "Gan Augmentation: Augmenting Training Data Using Generative Adversarial Networks", *arXiv Preprint arXiv:1810.10863*, 2018.

23. Mariani, G., F. Scheidegger, R. Istrate, C. Bekas and C. Malossi, "Bagan: Data Augmentation with Balancing Gan", *arXiv Preprint arXiv:1803.09655*, 2018.

24. Huang, S.-W., C.-T. Lin, S.-P. Chen, Y.-Y. Wu, P.-H. Hsu and S.-H. Lai, "Auggan: Cross Domain Adaptation with Gan-Based Data Augmentation", *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 718–731, 2018.

25. Lim, S. K., Y. Loo, N.-T. Tran, N.-M. Cheung, G. Roig and Y. Elovici, "Doping: Generative Data Augmentation for Unsupervised Anomaly Detection with Gan", *IEEE International Conference on Data Mining (ICDM)*, pp. 1122–1127, 2018.

26. Frid-Adar, M., I. Diamant, E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification", *Neurocomputing*, Vol. 321, pp. 321–331, 2018.

27. Kiyasseh, D., G. A. Tadesse, L. Thwaites, T. Zhu and D. Clifton, "Plethaugment: GAN-Based PPG Augmentation for Medical Diagnosis in Low-Resource Settings", *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 11, pp. 3226–3235, 2020.

28. Frid-Adar, M., E. Klang, M. Amitai, J. Goldberger and H. Greenspan, "Synthetic Data Augmentation Using GAN for Improved Liver Lesion Classification", *IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 289–293, 2018.

29. Madani, A., M. Moradi, A. Karargyris and T. Syeda-Mahmood, "Chest X-Ray Generation and Data Augmentation for Cardiovascular Abnormality Classification", E. D. Angelini and B. A. Landman (Editors), *Medical Imaging: Image Processing*, Vol. 10574, pp. 415 – 420, International Society for Optics and Photonics, SPIE, 2018.

30. Bhattacharya, D., S. Banerjee, S. Bhattacharya, B. U. Shankar and S. Mitra, "GAN-Based Novel Approach for Data Augmentation with Improved Disease Classification", *Advancement of Machine Intelligence in Interactive Medical Image Analysis*, pp. 229–239, Springer, 2020.

31. Xing, Y., Z. Ge, R. Zeng, D. Mahapatra, J. Seah, M. Law and T. Drummond, "Adversarial Pulmonary Pathology Translation for Pairwise Chest X-Ray Data Augmentation", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 757–765, Springer, 2019.

32. Kora Venu, S. and S. Ravula, "Evaluation of Deep Convolutional Generative Adversarial Networks for Data Augmentation of Chest X-Ray Images", *Future Inter-*

net, Vol. 13, No. 1, p. 8, 2021.

33. Ganesan, P., S. Rajaraman, R. Long, B. Ghoraani and S. Antani, "Assessment of Data Augmentation Strategies Toward Performance Improvement of Abnormality Classification in Chest Radiographs", *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 841–844, 2019.

34. Malygina, T., E. Ericheva and I. Drokin, "Data Augmentation with GAN: Improving Chest X-Ray Pathologies Prediction on Class-Imbalanced Cases", *International Conference on Analysis of Images, Social Networks and Texts*, pp. 321–334, Springer, 2019.

35. Mahapatra, D., "Generative Adversarial Networks and Domain Adaptation for Training Data Independent Image Registration", *arXiv preprint arXiv:1910.08593*, 2019.

36. Motamed, S., P. Rogalla and F. Khalvati, "RANDGAN: Randomized Generative Adversarial Network for Detection of COVID-19 in Chest X-Ray", *Scientific Reports*, Vol. 11, No. 1, pp. 1–10, 2021.

37. Segal, B., D. M. Rubin, G. Rubin and A. Pantanowitz, "Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs", *arXiv Preprint arXiv:2010.03975*, 2020.

38. Menon, S., J. Galita, D. Chapman, A. Gangopadhyay, J. Mangalagiri, P. Nguyen, Y. Yesha, Y. Yesha, B. Saboury and M. Morris, "Generating Realistic COVID19 X-Rays with a Mean Teacher + Transfer Learning GAN", *IEEE International Conference on Big Data (Big Data)*, pp. 1216–1225, 2020.

39. Khalifa, N. E. M., M. H. N. Taha, A. E. Hassanien and S. Elghamrawy, "Detection of Coronavirus (COVID-19) Associated Pneumonia Based on Generative Adversarial Networks and a Fine-Tuned Deep Transfer Learning Model Using Chest X-Ray

Dataset", *arXiv preprint arXiv:2004.01184*, 2020.

40. Kovalev, V. and S. Kazlouski, "Examining the Capability of GANs to Replace Real Biomedical Images in Classification Models Training", S. V. Ablameyko, V. V. Krasnoproshin and M. M. Lukashevich (Editors), *Pattern Recognition and Information Processing*, pp. 98–107, Springer International Publishing, Cham, 2019.

41. Salehinejad, H., S. Valaee, T. Dowdell, E. Colak and J. Barfett, "Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 990–994, 2018.

42. Chen, C., Q. Dou, H. Chen and P.-A. Heng, "Semantic-Aware Generative Adversarial Nets for Unsupervised Domain Adaptation in Chest X-Ray Segmentation", Y. Shi, H.-I. Suk and M. Liu (Editors), *Machine Learning in Medical Imaging*, pp. 143–151, Springer International Publishing, Cham, 2018.

43. Terzopoulos, D. and A.-A.-Z. Imran, "Multi-Adversarial Variational Autoencoder Networks", *18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 777–782, 2019.

44. Bigolin Lanfredi, R., J. D. Schroeder, C. Vachet and T. Tasdizen, "Interpretation of Disease Evidence for Medical Images Using Adversarial Deformation Fields", *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 738–748, Springer, 2020.

45. Deepshikha, K. and A. Naman, "Removing Class Imbalance Using Polarity-Gan: An Uncertainty Sampling Approach", *arXiv preprint arXiv:2012.04937*, 2020.

46. Choong, R. Z. J., S. A. Harding, B.-y. Tang and S.-w. Liao, "3-to-1 Pipeline: Restructuring Transfer Learning Pipelines for Medical Imaging Classification via Optimized GAN Synthetic Images", *42nd Annual International Conference of the*

IEEE Engineering in Medicine & Biology Society (EMBC), pp. 1596–1599, 2020.

47. Sundaram, S. and N. Hulkund, "GAN-based Data Augmentation for Chest X-Ray Classification", *arXiv Preprint arXiv:2107.02970*, 2021.

48. Arvanitidis, G., L. K. Hansen and S. Hauberg, "Latent Space Oddity: On the Curvature of Deep Generative Models", *International Conference on Learning Representations*, 2018.

49. Bojanowski, P., A. Joulin, D. Lopez-Pas and A. Szlam, "Optimizing the Latent Space of Generative Networks", J. Dy and A. Krause (Editors), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 600–609, PMLR, 10–15 Jul 2018.

50. Upchurch, P., J. Gardner, G. Pleiss, R. Pless, N. Snavely, K. Bala and K. Weinberger, "Deep Feature Interpolation for Image Content Changes", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7064–7073, 2017.

51. Antipov, G., M. Baccouche and J.-L. Dugelay, "Face Aging with Conditional Generative Adversarial Networks", *IEEE International Conference on Image Processing (ICIP)*, pp. 2089–2093, 2017.

52. Jahanian*, A., L. Chai* and P. Isola, "On the Steerability of Generative Adversarial Networks", *International Conference on Learning Representations*, 2020.

53. Goetschalckx, L., A. Andonian, A. Oliva and P. Isola, "Ganalyze: Toward Visual Definitions of Cognitive Image Properties", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5744–5753, 2019.

54. Shen, Y., J. Gu, X. Tang and B. Zhou, "Interpreting the Latent Space of Gans for Semantic Face Editing", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

55. Galatolo., F., M. Cimino. and G. Vaglini, "Generating Images from Caption and Vice Versa via CLIP-Guided Generative Latent Space Search", *Proceedings of the International Conference on Image Processing and Vision Engineering*, 2021.

56. Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin and J. Clark, "Learning Transferable Visual Models from Natural Language Supervision", *arXiv Preprint arXiv:2103.00020*, 2021.

57. Collins, E., R. Bala, B. Price and S. Susstrunk, "Editing in Style: Uncovering the Local Semantics of Gans", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5771–5780, 2020.

58. Härkönen, E., A. Hertzmann, J. Lehtinen and S. Paris, "GANSpace: Discovering Interpretable GAN Controls", H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (Editors), *Advances in Neural Information Processing Systems*, Vol. 33, pp. 9841–9850, Curran Associates, Inc., 2020.

59. Wu, Z., D. Lischinski and E. Shechtman, "StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12863–12872, June 2021.

60. Fetty, L., M. Bylund, P. Kuess, G. Heilemann, T. Nyholm, D. Georg and T. Löfstedt, "Latent Space Manipulation for High-Resolution Medical Image Synthesis via the StyleGAN", *Zeitschrift für Medizinische Physik*, Vol. 30, No. 4, pp. 305–314, 2020.

61. Fernández Blanco, R., P. Rosado, E. Vegas and F. Reverter, "Medical Image Editing in the Latent Space of Generative Adversarial Networks", *Intelligence-Based Medicine*, Vol. 5, p. 100040, 2021.

62. Desai, K. and J. Johnson, "Virtex: Learning Visual Representations from Textual

Annotations", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11162–11173, 2021.

63. Sariyildiz, M. B., J. Perez and D. Larlus, "Learning Visual Representations with Caption Annotations", A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm (Editors), *Computer Vision – ECCV*, pp. 153–170, Springer International Publishing, Cham, 2020.

64. Tan, H. and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers", *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

65. Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, 2019.

66. Su, W., X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei and J. Dai, "VL-BERT: Pre-Training of Generic Visual-Linguistic Representations", *International Conference on Learning Representations*, 2020.

67. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

68. Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold and S. Gelly, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *arXiv preprint arXiv:2010.11929*, 2020.

69. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, "Attention is All you Need", I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Editors), *Advances in Neural Information Processing Systems*, Vol. 30, p. 5998–6008, Curran Associates, Inc., 2017.

70. Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, "Chestx-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106, 2017.

71. Johnson, A. E., T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, "MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports", *Scientific Data*, Vol. 6, No. 1, pp. 1–8, 2019.

72. Ye, W., J. Yao, H. Xue and Y. Li, "Weakly Supervised Lesion Localization with Probabilistic-CAM Pooling", *arXiv preprint arXiv:2005.14480*, 2020.

73. Neumann, M., D. King, I. Beltagy and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing", *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327, Association for Computational Linguistics, Florence, Italy, Aug. 2019.

74. Honnibal, M., I. Montani, S. Van Landeghem and A. Boyd, *spaCy: Industrial-Strength Natural Language Processing in Python*, 2020.

75. Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Editors), *Advances in Neural Information Processing Systems*, Vol. 30,

pp. 6626–6637, Curran Associates, Inc., 2017.

76. Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

77. Brownlee, J., *ROC Curves and Precision-recall Curves for Imbalanced Classification*, 2020, `https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/`, accessed in March 2022.

78. Hu, J., J. Li, Z. Chen, Y. Shen, Y. Song, X. Wan and T.-H. Chang, "Word Graph Guided Summarization for Radiology Findings", *arXiv Preprint arXiv:2112.09925*, 2021.

79. Yuan, Z., Y. Yan, M. Sonka and T. Yang, "Large-Scale Robust Deep Auc Maximization: A New Surrogate Loss and Empirical Studies on Medical Image Classification", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3040–3049, 2021.

APPENDIX A: LEGAL NOTICE

The images that emerged within the scope of this thesis work and whose copyrights were transferred to the publisher were used in the thesis book in accordance with the publishing policy valid for the reuse of the text and graphics produced by the author on the website of the publisher.