XRAY-GAN: Conditional and Unconditional Score-guided GAN for Chest X-ray Pathology Classification

Joy Chen [1] **Qingfu Wan** [1]

1. Introduction

The popularization of deep learning models in the past few years has extended to sectors from self-driving cars to election predictions to music composition. A deep learning utilization of particular note and importance is within the healthcare industry. Medical imaging requires radiology technicians, lab space, consenting patients, and representative samples. Teaching a computer to classify x-rays accurately could corroborate existing diagnoses, track changes in patients over time, and even extend to more complex medical datasets such as brain MRIs.

A handful of labeled thoracic x-ray datasets have been released over the years. Although relatively large and well-documented, machine learning researchers have not yet found a reliable way to train classifiers on this data. The difficulty of condition classification lies in the comprehensiveness of the image sets and the minute complexity of any such x-ray diagnosis. A generative adversarial network (GAN) (Goodfellow et al., 2014) outperforms other deep generative models by sidestepping complicated probability calculations. We propose and compare two GAN methods to augment existing chest x-ray datasets during classification training.

Code is at `https://github.com/strawberryfg/xraygan`.

2. Related Works

2.1. Perceptual Judgment of Image Generation

Assessing the quality and diversity of machine-generated images is inherently hard. One way is to use manual perceptual judgment. For instance, generated chest X-rays were evaluated by radiologists in (Salehinejad et al., 2018b). (Elgammal et al., 2017) required user studies on GAN-generated artworks. However, human intervention is costly and will lead to biased estimates. To automate the evaluation, (Salimans et al., 2016) instead stated that **Inception Score** correlates well with human evaluation. In the realm of *Style Transfer* where the task is to generate an image that encompasses the style and content of the target images, *perceptual loss functions* are usually used to automatically measure statistical differences between the generated image and the target image (Johnson et al., 2016). *We made use of **Inception Score** and perceptual losses as auxiliary objectives to guide the GAN (Sec. 4.2).*

2.2. Uncertainty Sampling

Uncertainty Sampling is a strategy for identifying unlabeled items that are near a decision boundary in the current model. (Munro, 2020) One approach, *least confidence sampling* gauges the difference between the most confident prediction and 100% prediction, thus capturing how confident (uncertain) that prediction is. In this regime, pseudo-labeling (Lee et al.) regards the class from the maximum predicted probability as the true label. *We borrowed the idea of pseudo-labeling and adapt it to annotate unlabeled GAN-generated chest X-ray images. (Sec. 4.2)*

3. Preliminaries

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) consist of two models: a generator G and a discriminator D. G learns from the real data and generates sample images while D evaluates the probability that a sample is real or generated by G. Alternately optimizing these two objective functions allows the generated samples to improve without having to evaluate complex probabilistic calculations. The two-player min-max game optimizes the value function V with data $\mathbf{x}$. The prior input noise is $p_z(z)$ and y has hidden representation in G.

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim p_{\mathbf{x}}}[logD(\mathbf{x})] + E_{\mathbf{z}\sim p_{\mathbf{z}}}[log(1-D(G(\mathbf{z})))] \quad (1)$$

.

Evaluating GANs Ideal models will have well-defined bounds and produce diverse samples which agree with human judgement (Borji, 2018). In our work, we focus mainly on diversity. GANs are prone to issues such as mode col-

[1]Courant Institute of Mathematical Sciences, New York University, New York, USA.

NYU CSCI-GA 2565-001 Machine Learning Spring 2021

lapse and over-fitting, which can be prevented by prioritizing generated image diversity. A good metric will reward models with a high entropy marginal label distribution $\int_x p(|x)p_g(x)$. Other important features include discriminability and robustness.

Inception Score (**IS**) is the most widely-used GAN evaluation metric. **IS** takes the output of the Inception v3 network on generated data, from which it computes the statistics.

$$\begin{aligned}\mathcal{IS}(\mathcal{G}) &= \exp(E_{\mathbf{z}}[KL(P_g(y|\mathcal{G}(\mathbf{z}))||P_g(y))]) \\ &= \exp(H(y) - E_{\mathbf{z}}[y|\mathcal{G}(\mathbf{z})])\end{aligned} \tag{2}$$

Eq. 2 expects meaningful generated data (low entropy of $P_g(y|\mathcal{G}(\mathbf{z}))$) and varied generation (high entropy of $P_g(y)$). In practice, empirical marginal class distribution $\tilde{P}_g(y) = \frac{1}{N}\sum_{i=1}^{N} P_g(y|\mathcal{G}(\mathbf{z}^{(i)}))$ is used. The approximation is then:

$$\tilde{\mathcal{IS}}(\mathcal{G}) = \exp(\frac{1}{N}\sum_{i=1}^{N} KL(P_g(y|\mathcal{G}(\mathbf{z}^{(i)}))||\tilde{P}_g(y))) \tag{3}$$

Direct training with **IS** will overfit to Inception (Lee & Seok, 2020), we circumvent this using our own classifier.

Fréchet Inception Distance (FID) finds the distance between the embedded Gaussians produced by generated samples in an intermediate layer of Inception-v3. While **IS** evaluates only the generated images, **FID** takes into account both the generated image distributions and the real training data. r, g are the real and generated embeddings, respectively, and μ, Σ are the mean and covariance of the specified distribution.

$$FID(r,g) = ||\mu_r - \mu_g||_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{\frac{1}{2}}) \tag{4}$$

IS vs FID: **IS** estimates the diversity of the generated images while **FID** measures the distance between the generated and real distributions. In that regard, **FID** is more sensitive to mode collapse. **FID** has been shown to align more closely to human judgment as well as outperform **IS** with noise level (Heusel et al., 2018). However, **IS** outperforms **FID** as multiple artifacts are added to the image (Borji, 2018). We opt to **IS** for training; **FID** for evaluation.

Maximum Mean Discrepancy (**MMD**) is a measurement of dissimilarity between two probability distributions P_r and P_g using samples independently drawn from each. **Kernel MMD** uses a fixed kernel k, *e.g.* the Gaussian kernel $k(\mathbf{x}, \mathbf{x}') = \exp(||\mathbf{x} - \mathbf{x}'||^2)$ in this paper. We fed $P_r(P_g)$ with class distributions on real(fake) data during training.

$$\begin{aligned}M_k(P_r, P_g) = E_{\mathbf{x},\mathbf{x}'\sim P_r}[k(\mathbf{x},\mathbf{x}')] \\ - 2E_{\mathbf{x}\sim P_r,\mathbf{y}\sim P_g}[k(\mathbf{x},\mathbf{y})] + E_{\mathbf{y},\mathbf{y}'\sim P_g}[k(\mathbf{y},\mathbf{y}')]\end{aligned} \tag{5}$$

Style Gram Matrix In style transfer, *Gram Matrix* is employed to penalize style differences between the generated and the style image (Johnson et al., 2016). Each entry of the *Gram Matrix* $G_j^{\phi}(x) \in R^{C_j \times C_j}$ is $G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j \times H_j \times W_j}\phi_j^c(x)^T\phi_j^{c'}(x)$ measuring the unnormalized *cosine similarity* between 2 feature vectors at channel c and c', which will be high if the feature maps $\phi_j^c(x), \phi_j^{c'}(x)$ correlate well. The *style reconstruction loss* is then the squared Frobenius norm of the difference between the Gram matrices of the generated $\hat{I}$ and the style target I: $\ell_{style}^{\phi,j}(\hat{I}, I) = ||G_j^{\phi}(\hat{I}) - G_j^{\phi}(I)||_F^2$. Our *style loss* is based on this matrix.

4. Method

We present two approaches to generate chest X-ray images for pathology classification.

The first approach conditional GAN (**CGAN**) conditions upon class labels in an attempt to equalize the imbalanced multi-modal data.

The second approach unconditional score-guided GAN (**UCGAN**) is not conditioned on class labels.

4.1. Conditional Generative Adversarial Network

Conditional GAN (**CGAN**) (Mirza & Osindero, 2014) extends basic GAN by conditioning with a layer of auxiliary information y. Unlike vanilla GAN, this allows the user to control the modes and hopefully ascertain characteristics of specific types to the generated images. Conditioning upon labels logically follows within a diagnostic setting. The new min-max game is simply the original GAN formula with x and z conditioned upon y.

$$\begin{aligned}\min_G \max_D V(D,G) = E_{\mathbf{x}\sim p_{\mathbf{x}}}[logD(\mathbf{x}|\mathbf{y})] \\ + E_{\mathbf{z}\sim p_{\mathbf{z}}}[log(1 - D(G(\mathbf{z}|\mathbf{y})))]\end{aligned} \tag{6}$$

4.2. Unconditional Generative Adversarial Network

Without *a priori* class label information, in this part we elaborate on the unconditional GAN guided by 3 scores:

- **Inception Score Loss**: We replaced Inception in **IS** (Eq. 2) that is trained on ImageNet with our readily available task-specific classifier. The exponential term is also discarded for simplicity.

$$\begin{aligned}\mathcal{L}_{IS}(\mathbf{z};\mathcal{G},\mathcal{C}) &= E_{\mathbf{z}}[KL(P_g(y|\mathcal{G}(\mathbf{z}))||\tilde{P}_g(y))] \\ &= \tilde{H}(y) - E_{\mathbf{z}}[H(y|\mathcal{G}(\mathbf{z}))]\end{aligned} \tag{7}$$

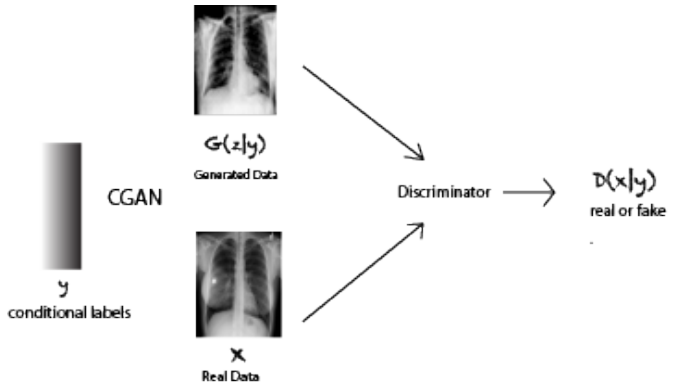

Figure 1. **Diagram of the conditional GAN**.

$\tilde{P}_g(y)$ is the empirical marginal distribution $\frac{1}{N}\sum_{i=1}^{N} P_g(y|\mathcal{G}(\mathbf{z}^{(i)}))$. ($N$ is batch size) Intuitively, Eq. 7 favors $\tilde{P}_g(y)$ to have high entropy meaning diverse generated samples within each batch, and $P_g(y|\mathcal{G}(\mathbf{z}))$ to have low entropy meaning correct classification on generated samples. Rather than an approximation of the true **Inception Score**, this score can be interpreted as an entropy regularization term.

- **Maximum Mean Discrepancy Loss**: We used finite samples within the batch from P_r and P_g to estimate **MMD** distance. Given $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \sim P_r$, $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \sim P_g$ [1], the estimator of $M_k(P_r, P_g)$ is

$$\tilde{M}_k(X,Y) = \frac{1}{\binom{N}{2}}\sum_{i\neq i'} k(\mathbf{x}_i, \mathbf{x}_i') - \frac{2}{\binom{N}{2}}\sum_{i\neq j} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{\binom{N}{2}}\sum_{j\neq j'} k(\mathbf{y}_j, \mathbf{y}_{j'}) \quad (8)$$

, which is the finite sample approximation of the expectation. A lower **MMD** means P_g is closer to P_r, indicating the classifier produces a similar probability distribution for both real (P_r) and generated data (P_g).

- **Style Gram Matrix Loss**: Write I_r as the real image (*style target*), I_g as the generated image (*content target*) by **UCGAN**, $\hat{I}$ as the style transferred image. The original *style reconstruction loss* is $\ell^{\phi,j}_{style}(\hat{I}, I_r) = ||G^\phi_j(\hat{I}) - G^\phi_j(I_r)||^2_F$. We insert another *Style Loss*:

$$\ell^{\phi,j}_{style}(I_g, I_r) = ||G^\phi_j(I_g) - G^\phi_j(I_r)||^2_F \quad (9)$$

Assuming similar style ($e.g.$ common colors, textures, and patterns) indicates similar inter-channel correlation, a low *Style Gram Matrix Loss* in Eq. 9 reveals a high style similarity between generated X-rays and real X-rays within each batch.

The total regularization score is [2]

$$\mathcal{L}_{reg} = \alpha\mathcal{L}_{IS}(\mathbf{z};\mathcal{G},\mathcal{C}) + \beta\tilde{M}_k(P_r(\mathbf{x}), P_g(\mathcal{G}(\mathbf{z}))) + \gamma(\ell^\phi_{style}(\hat{I}, I_r) + \ell^\phi_{content}(\hat{I}, I_g)) + \eta\ell^\phi_{style}(I_r, I_g) \quad (10)$$

With GAN losses and classification losses, the total loss is

$$\mathcal{L}_{total} = \mathcal{L}_{reg} + \delta\mathcal{L}_G(\mathbf{z}) + \lambda\mathcal{L}_D(\mathbf{x},\mathbf{z}) + \tau\mathcal{L}_\mathcal{C}(\mathbf{x},\mathbf{z}) \quad (11)$$

Pseudo-labels To obtain labels of generated data, we used the pseudo-labeling method (Lee et al.) which assigns a label based on the most likely class predicted by the current $\mathcal{C}$. We only kept those with a predicted probability above a threshold t (Sec. 5.2.1), which are then fed to the classifier immediately in each mini-batch. The loss of generated data is scaled by an adversarial weight ξ, which controls the relative importance of generated data versus real data.

5. Experiments

Dataset We used the NIH ChestX-ray 14 dataset (Wang et al., 2017) containing 112,120 frontal-view chest X-ray images of 30,805 patients. The data spans 14 finding labels, along with information on patient age and gender.

We also conducted preliminary work with the Labeled Chest X-ray Images provided by the University of California San Diego (Kermany et al., 2018). The thousands of x-rays within this dataset are focused solely on pneumonia diagnosis with a "normal" set as well as a "pneumonia" set.

Metrics

[1] P_g/P_r is the class distribution rather than the real/generated data distribution. In **MMD** X and Y are not images and labels.

[2] We summed up the 2 *Style Gram Matrix Losses* and the original *content reconstruction loss* for all layers j.

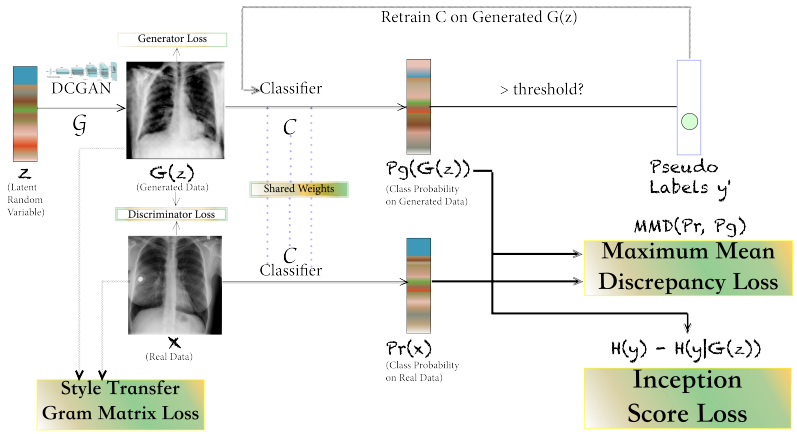

Figure 2. **Diagram of the unconditional score-guided GAN. UCGAN** is trained on generated and real images simultaneously with a shared classifier C. We used *pseudo labeling* that automatically labels generated images whose current maximum class predictive probability is above a threshold. We added 3 regularization terms. **Inception Score Loss** encourages generated images to be diverse and to be highly classifiable. **Maximum Mean Discrepancy Loss** matches class probability distribution of generated images with that of real images. **Style Transfer Gram Matrix Loss** retains the style (color, texture, and common patterns) in real data during generation.

- *AUC:* The ROC curve is the true positive rate (TPR) as a function of the false positive rate (FPR) at various threshold settings. AUC is simply the area under this curve. The higher AUC is, the better the classifier is.
- *Precision/Recall:* The confusion matrix (Tab. 1) encapsulates the classification options at the intersection of real and predicted, with true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The classic accuracy metric gives the total correct predictions over all predictions. However, in certain cases, the preferred methods prioritize different aspects of the confusion matrix.

		Real	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Table 1. **Confusion matrix.**

Precision P is the ratio of correctly predicted positives to total predicted positives.

$$P = \frac{TP}{TP + FP} \tag{12}$$

Recall R refers to the ratio of correctly predicted positives to total actual positives.

$$P = \frac{TP}{TP + FN} \tag{13}$$

- *Note on evaluation metric choice:* The most effective (or least damaging) evaluation metric must consider the implications of error. Some of the lung conditions in question hold immense gravity in a person's life. Would we rather have a false positive or negative? Nonetheless, it is imperative in matters of people's lives that we use the results of machine learning to augment existing systems rather than rely on them in their infancy.

Class Imbalance Chest x-ray14 has a strong class imbalance across the 14 labels. We conditioned by label in **CGAN** to account for this imbalance and generalized to a few difficult labels rather than whole. For the conditional GAN, we used Pneumonia, Effusion, Mass, and Hernia. For the unconditional score-guided GAN, we used easily trainable classes Pneumonia, Edema and Hernia.

Baseline To establish a baseline for **UCGAN**, we trained a vanilla ResNet-50. Note only real samples are used, which does not lead to satisfactory results as evident in Tab. 2.

	Pneumonia	Edema	Hernia
AUC	0.63	0.75	0.88

Table 2. **AUC of a vanilla ResNet-50 on Chest X-ray 14.**

5.1. CGAN

In our experiments with **CGAN**, we trained on the labels Pneumonia, Effusion, Mass, and Hernia. We received an FID-score for the model of 6.221 using PyTorch FID (Seitzer, 2020), where a lower score is better. The resultant

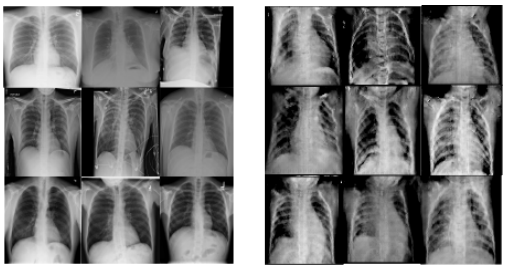

Figure 3. **Left: real Chest X-ray14 images. Right: Fake x-rays generated by CGAN.**

images are compared to the real images in Fig. 3. There is a clear indication of ribs, spine, and heart shading, with less defined lung areas. The high-contrast bone structure takes away from the identifiable variations of the lungs. We then trained a classifier on the generated images and assessed the performance with precision and recall, as shown in Tab. 3.

	Pneumonia	Effusion	Mass	Hernia
Precision	0.397	**0.639**	**0.430**	0.592
Recall	**0.428**	0.581	0.326	0.587

Table 3. **Precision and recall values.** The results from classifying **CGAN** generated images trained on Chest X-ray 14.

5.2. UCGAN

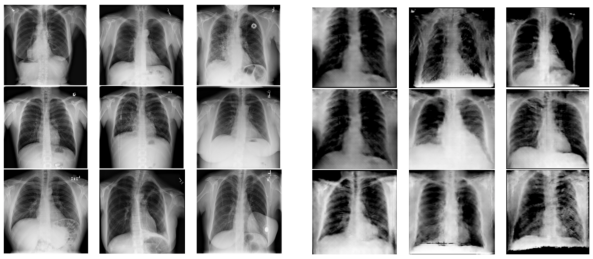

Figure 4. **Samples of real and synthetic images by UCGAN**.

Without balancing among classes, we trained **UCGAN** on Pneumonia, Edema, and Hernia. In Fig. 4, we present real and artificial X-rays on the NIH dataset. Some parts of the generated ribs are unrecognizable, whereas the gradient and contrast resemble real X-rays better than **CGAN**. Next, we perform ablative studies.

5.2.1. Pseudo-label Threshold

A high threshold (low uncertainty score for *least confidence sampling*) limits the amount of generated data used by $\mathcal{C}$, while a low threshold complicates the classification (high uncertainty score). We empirically found the optimal t to be around 0.8, as shown in Tab. 4.

Threshold t	Pneumonia	Edema	Hernia
0.6	0.57	0.82	**0.94**
0.7	0.56	0.82	0.93
0.8	**0.69**	**0.82**	0.90
0.9	0.54	0.81	0.92

Table 4. **AUC against t for pseudo-labels in UCGAN.**

5.2.2. Losses

The importance of additional scores can be understood as follows: The **Style Gram Matrix Losses** conform the style of generated data to that of real data. *Implicitly* guided by the classifier, **Inception Score Loss** and **Maximum Mean Discrepancy Loss** align the distribution of generated data to that of real data. Furthermore, **Inception Score Loss** prevents **UCGAN** from *mode collapse* by regularizing the entropy. A combination of these 3 is generally the best, as displayed in Tab. 5.

Method	Pneumonia	Edema	Hernia
Simple	0.69	0.78	0.82
+ **IS**	0.68	0.80	0.91
+ **IS** + **MMD**	0.67	0.82	**0.94**
+ **IS** + **MMD** + **Gram**	**0.70**	**0.83**	0.92

Table 5. **Effectiveness of the scores in UCGAN.** Simple: w/o $\mathcal{L}_{reg}$ in Eq. 11. **Gram**: *Style Loss* in Eq. 9. AUC is applied.

5.2.3. Requirement of Number of Real Samples

Data and labels are often scarce because of lacking resources. To test an extreme case, we used few real samples. Specifically, we trained a binary classification system to detect Edema guided by Eq. 11 with a varying number of real samples. On measuring AUC in Tab. 6, it is interesting to note that as few as 40 real samples are enough for **UCGAN** to learn the data distribution of Edema, which aligns with the finding in (Haque, 2020).

Number of Real Samples	AUC
0	0.53
20	0.52
30	0.62
40	**0.71**
100	**0.71**

Table 6. **AUC against real Edema samples useds in UCGAN.**

5.3. Comparison with State-of-the-art Works

5.3.1. CGAN

Precision/Recall on X-ray 14 Effusion Our results from **CGAN** on the x-ray sets were comparable with those from the original code by Gupta and Lynch (2019), but we chose more unbalanced labels. We compare our **CGAN** implementation on 4 labels with a proposed balanced DCGAN

(Salehinejad et al., 2018a) run on 5 labels, with the common label being 'Effusion'. The common metric is precision and recall. While we assessed a classifier on only generated images, their "Imbalanced Real" dataset consists of both real and generated images. As seen in Tab. 7, our precision/recall values are lower (within 0.15) which could be attributed to their inclusion of real data in classification. The focus of this iteration was to assess the quality of the generated images. As our goal was to use generated images to supplement existing datasets, further work should be done on the combination of our **CGAN**-generated images and the training data.

	Precision	Recall
(Salehinejad et al., 2018a)	0.68	0.72
Our CGAN	0.639	0.581

Table 7. **CGAN** on X-ray 14 comparison with previous work.

Accuracy on Zhang Lab Data Pneumonia We compare our **CGAN** implementation on the Zhang Lab Data (Kermany et al., 2018) with results from the proposed ECGAN (Haque, 2021). In this case, we classified with both generated and real data, as did ECGAN. The shared metric is accuracy. As seen in Tab. 8, our accuracy was on par with other state-of-the-art GAN iterations.

ECGAN (Haque, 2021)	90.875 $\pm$2.245
Our CGAN	89.13 $\pm$1.22

Table 8. **CGAN** on Zhang data comparison with previous work

5.3.2. UCGAN

AUC on X-ray 14 Pneumonia & Edema & Hernia We examine the state-of-the-art performance of **UCGAN** on the 3 selected classes in Tab. 9. As shown, **UCGAN** achieves the best result on Hernia. On the other two, our performance is on par with leading methods. We speculate that the gap probably arises from **(1)** imbalanced classes **(2)** under-trained backbone ResNet-50 for $\mathcal{C}$.

Method	Pneumonia	Edema	Hernia
Li (Li et al., 2018)	0.63	0.71	0.67
CRAL (Guan & Huang, 2020)	0.73	**0.85**	0.92
MT (Liu et al., 2021)	**0.74**	0.85	0.83
CheXNet (Rajpurkar et al., 2017)	0.70	0.84	0.89
Baseline: ResNet-50	0.63	0.75	0.88
Our UCGAN	0.70	0.83	**0.92**

Table 9. **UCGAN** on X-ray 14 AUC comparison with prior works.

5.4. Implementation Details

- **CGAN** We implemented **CGAN** based on the original paper (Mirza & Osindero, 2014)) and a later iteration for x-rays by Gupta and Lynch (https://github.com/Anushka1610/chest-gan/blob/master/model/cgan.py). We used Adam with a learning rate of 0.0002 and first moment decay rate of 0.5. We used both ReLU and sigmoid in the discriminator, but only ReLU in the generator.

- **UCGAN** α, β, γ, η, δ, λ, τ in Eq. 10, Eq. 11 are -0.25, 3, 0.001, 0.015, 1.0, 0.5 and 1.0 respectively. The adversarial weight ξ is 0.125. The corresponding base learning rates for $\mathcal{G}$, $\mathcal{D}$, $\mathcal{C}$ are 0.0005, 0.001, and 0.0005.[3] We used Adam with L2 weight decay of 0.001 and a batch size of 12. We set the threshold t to 0.75. We used DCGAN (Radford et al., 2015) for $\mathcal{G}$ and $\mathcal{D}$[4]. We used a VGG loss network for style transfer following (Johnson et al., 2016). The backbone of $\mathcal{C}$ is ResNet-50, and cross entropy loss is adopted.

6. Discussion

We saw the early stage effectiveness of both unconditional and conditional GANs for thoracic x-ray generation. The work in this paper can extend with more experimentation in classifiers, a stronger backbone, and combining conditioning on both real labels and pseudo-labels with **UCGAN**'s loss properties.

Apart from that, we plan to incorporate techniques from few-shot learning, unsupervised learning, and uncertainty sampling into a unified semi-supervised learning algorithm. We will also devise new probability distance metrics for image generation and classification. More generally, generative models for x-rays could benefit from including non-image features such as age and gender.

7. Conclusion

CGAN finds success in separating modes by label and harnessing the characteristics essential to classification. Our implementation struggles the minuscule nuances in lung gradation that can identify one or more diseases.

The scores in **UCGAN** are conducive to learning the real data distribution by *explicitly* aligning with human perceptual judgment. We have shown that generated images by **CGAN** already benefit the downstream classification. Further, the integration of a classification head into **UCGAN** *implicitly* regularizes the generation for practical usage.

[3] $\mathcal{G}$ was updated 3 times per update of $\mathcal{D}$ at first, followed by a flip of the gradient update rate. Finally, the rate was $1:1$. We implemented in PyTorch on an 8GB NVIDIA GTX 1070 card.

[4] The generated I_g is 128×128, the real I_r is resized to 128×128 from 1024×1024 before feeding to $\mathcal{D}$ and $\mathcal{C}$.

References

Borji, A. Pros and cons of gan evaluation measures. *arXiv preprint arXiv:arXiv:1802.03446*, 2018.

Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Guan, Q. and Huang, Y. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 130:259–266, 2020.

Haque, A. EC-GAN: low-sample classification using semi-supervised algorithms and gans. *CoRR*, abs/2012.15864, 2020. URL https://arxiv.org/abs/2012.15864.

Haque, A. Ec-gan: Low-sample classification using semi-supervised algorithms and gans. *arXiv preprint arXiv:2012.15864*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Kermany, D., Zhang, K., and Goldbaum, M. Large dataset of labeled optical coherence tomography (oct) and chest x-ray images, 2018.

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.

Lee, M. and Seok, J. Score-guided generative adversarial networks. *arXiv preprint arXiv:2004.04396*, 2020.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., and Fei-Fei, L. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Liu, F., Tian, Y., Cordeiro, F. R., Belagiannis, V., Reid, I., and Carneiro, G. Self-supervised mean teacher for semi-supervised chest x-ray classification. *arXiv preprint arXiv:2103.03629*, 2021.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:arXiv:1411.1784*, 2014.

Munro, R. Human-in-the-loop machine learning. *Sl: O'REILLY MEDIA*, 2020.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

Salehinejad, H., Colak, E., Dowdell, T., Barfett, J., and Valaee, S. Synthesizing chest x-ray pathology for training deep convolutional neural networks. *IEEE Transactions on Medical Imaging*, PP:1–1, 11 2018a. doi: 10.1109/TMI.2018.2881415.

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., and Barfett, J. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 990–994. IEEE, 2018b.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*, 2016.

Seitzer, M. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.