Latent-optimization based Disease-aware Image Editing for Medical Image Augmentation

Aakash Saboo[1]
aakashsaboo2@gmail.com
Prashnna K. Gyawali[2]
pkg2182@rit.edu
Ankit Shukla[3]
e20soe809@bennett.edu.in
Neeraj Jain[4]
jainneeraj@hotmail.com
Manoj Sharma[3]
manoj.sharma1@bennett.edu.in
Linwei Wang[2]
lxwast@rit.edu

[1] Delhi Technological University
New Delhi,India

[2] Rochester Institute of Technology
Rochester
NY, United States

[3] Bennett University
UP, India

[4] Sir Ganga Ram Hospital
New Delhi, India

Abstract

Data augmentation addresses the critical challenge of limited data in medical imaging. While generative adversarial networks (GANs) have been a popular choice in synthesizing medical images, controlled generation targeting disease-specific semantic has been difficult, partly due to the difficulty to disentangle local disease-specific semantic factors from global disease-irrelevant factors. In this work, we present a semantic image editing framework for medical image augmentation that is able to generate smooth variations along the desired direction of disease attributes in user-defined regions of interest. This is achieved by discovering the optimal trajectory on the latent manifold of a pre-trained StyleGAN, guided by a mask of the region of interest and explicitly constrained by desired directions of semantic changes. We test the presented method on the public Chest X-ray dataset. To evaluate the quality of the generated medical images, we leverage both domain experts (pulmonologists) for qualitative assessments and present a novel metric to quantify the ability of the presented method to generate progression of disease severity in the synthesized images. Finally, we demonstrate that data augmentation using the presented method improves generalization for downstream classification tasks.

1 Introduction

Limited labeled dataset is one of the fundamental challenge for generalization in medical imaging [10] [23]. Data augmentation is a standard approach to increase the sample size for achieving generalizable models. While generative adversarial networks (GANs) have been a popular choice in synthesizing medical images [3, 5, 6, 14, 15, 18, 19, 21, 24], controlled generation targeting disease-specific semantic has been difficult.

Outside medical image augmentation, GAN-based semantic image editing techniques have been presented to provide better control over the generated samples by, for instance, traversing over disentangled latent space [8, 20, 26]. Most existing approaches, however, are *global*: the generated variations can potentially affect all the pixels in the image, unless the latent direction for traversing is perfectly disentangled to a specific semantic factor/region of interest. The latter is unfortunately challenging in medical images. As a result, while successful in application domains of creativity and design, GAN-based image editing has been little considered for medical image augmentation. In [10], disentangled latent codes obtained by variational autoencoders (VAEs) were exploited to manipulate physical attributes such as torso rotation and lobe size in X-ray images of the lung. The generated images by VAE however were low in resolution, and disease-specific semantic factors were reported to be more difficult to disentangle in comparison to global disease-irrelevant factors (*e.g.*, global torso variations). In [17], synthetic X-ray images were generated by linear latent space traversal along a semantic direction but only for cardiac silhouette manipulation.

Most recently, generating only *local* variations was shown possible by optimization over the latent manifold of GAN guided by a mask of region of interest in the image space [26]. This provides an excellent candidate to augment medical images incorporating high-level domain knowledge of disease-related features or anatomy within the image. The current approach, however, is not able to discover directions for traversal that can result in smooth semantic changes in the image space. This is fundamentally because low-dimensional manifold learned by GAN is non-linearly related to the image space [4]. As a result, unconstrained optimization in the latent space will introduce variations in image samples in all possible semantic directions, even when guided by a mask of region of interest.

In this paper, we present a semantic image editing framework for medical image augmentation that is able to generate smooth variations along the desired direction of disease attributes in user-defined regions of interest in medical images. This is achieved by discovering the optimal trajectory on the latent manifold of a pre-trained StyleGAN, guided by a mask of region of interest and explicitly constrained by desired directions of semantic changes specific to task-related variations. For the semantic directions, we choose the principal components that results in maximum variations along the given disease-related features. We then use this disease-specific semantic direction to guide the search for the optimal trajectory on the latent manifold of a pre-trained GAN, which achieves smooth and monotonous changes in the desired disease-related image attributes. To evaluate the quality of the generated medical images, we leverage both domain experts (pulmonologists) for qualitative assessments, and present a novel metric to quantify the ability of the presented method to generate progression of disease severity in the synthesized images. Finally, we demonstrate that data augmentation using the presented method improves generalization for downstream tasks (*e.g.*, classification). We test the presented method on the public Chest X-ray dataset [13], and focus our analyses on two super-class disease categories of cardiomegaly and lung opacity. To summarize, the main contributions of this work include:

- We present a semantic image editing framework for producing disease-aware variations in user-defined local region of interest.
- We present an approach to optimize a trajectory on the latent manifold by explicitly controlling the latent-optimization via the semantic directions of task-related features.
- We evaluate the clinical validity the generated samples both with the aid of domain

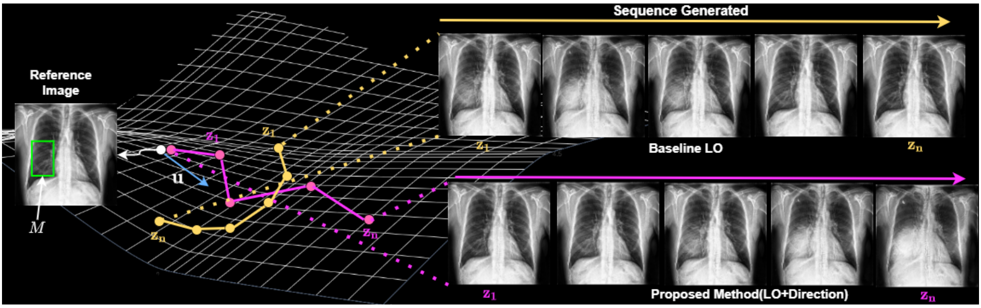


Figure 1: Schematic diagram of a semantic image editing framework via latent-optimization in a user-defined local region of interest (green bounding box). The baseline latent optimization method generates images (top sequence) with non-smooth changes. The presented method generates images (bottom sequence) with smooth and monotonous changes along the desired semantic directions (denoted by **u**).

> experts (pulmonologists) and via a novel quantitative metric to assess the ability of the presented method to generate progression of disease severity within a user-defined region.

To our knowledge, this work is the first to generate medical image samples with disease-aware and localized variations for improving a downstream task.

2 Methods

We begin with a generative model StyleGAN [22], denoted as G, to generate state-of-the-art high-resolution images and to offer controllable and editable latent features [1].

2.1 Disease-Aware Latent Optimization Framework

We first start with latent optimization for a given generated image $\mathbf{x}$ (hereafter referred to as reference image) within a user-defined localized region. For $\mathbf{x}$, we obtain latent vector $\mathbf{z}$ in $\mathcal{W}+$ space for StyleGAN architecture, and use a subset of it for optimization purpose. In this framework, our objective is to find optimal points $\mathbf{z}_i, \forall i \in \{1 \ldots n\}$ in the latent space of G such that the generated image $\mathbf{x}_i = G(\mathbf{z}_i), \forall i \in \{1 \ldots n\}$ exhibit variations within a user-defined rectangular masked region M over $\mathbf{x}$.

Let us define $\mathbf{x}_M$ as the image formed by cropping the masked region and $\mathbf{x}_M^*$ as its complement image (*i.e.*, unmasked region). Following previous work [26], our optimization framework for each $\mathbf{x}_i = G(\mathbf{z}_i)$ first comprises of image-space based objective:

$$\mathcal{L}_X\left(\mathbf{x}, \mathbf{x}_i; M\right) = \left|D\left(\mathbf{x}_M, \mathbf{x}_{i,M}\right) - c\right| + D\left(\mathbf{x}_M^*, \mathbf{x}_{i,M}^*\right) \tag{1}$$

where D can be any pixel-wise distance (e.g., $L1$, $L2$ or SSIM [25]). This objective helps us find the latent points with up to c units of variation in the masked region through the first term, while conserving the unmasked part through the second term. We then consider latent-space based objective where we apply spring loss to the discovered points $\mathbf{z}_i$ in order create smoothly varying image samples:

$$\mathcal{L}_{\text{spring}}\left(Z;k\right)=\sum_{i=1}^{n-k}\left(\|\mathbf{z}_i-\mathbf{z}_{i+k}\|_2-k\sigma\right)^2 \tag{2}$$

where σ is the rest length of springs between each vector in series, encouraging smooth variation and k represents length of the series. Combining variants of Eq.1 and Eq. 2, for a given reference latent vector $\mathbf{z}$, and masked region M, we obtain following optimization objective:

$$\tilde{Z}=\underset{Z}{\arg\min}\quad \alpha\sum_{i=1}^{n}\mathcal{L}_X\left(\mathbf{x},\mathbf{x}_i;M\right)+\beta\mathcal{L}_{\text{spring}}\left(Z;1\right)+\gamma\mathcal{L}_{\text{spring}}\left(Z;2\right) \tag{3}$$

where α,β,γ are the hyper-parameters. Optimization of objective function in Eq. 3 leads to discovery of latent points with irregular variations in the masked region (demonstrated later in the experiments) due to lack of information about the direction of the latent trajectory. Therefore, in order to have controlled variations, we propose to include the direction of the disease-specific variations in the optimization process.

Disease-Specific Semantic Directions: We identify important latent directions based on Principal Component Analysis (PCA) applied on the latent representation of generated images to control latent-optimization semantics. We manually analyze each of the principal components (PC) individually, and choose the PC that results in maximum variations along the given disease-related features (*e.g.,* increase in heart size for cardiomegaly). We denote this direction as $\mathbf{u}$. These principal components span the major variations expected in the medical images, and often the latent factors are entangled together. This phenomenon has been observed in the previous work, where principal components from StyleGAN have resulted in the entanglement of facial attributes like gender and head rotation [12]. Although it is desirable to have perfect disentangled directions, in our case we will show that it is not necessary because these directions work only as a constraint during latent optimization.

Directional Latent Optimization: In order to guide the latent trajectory, we assume that proper samples (which incur changes only in the masked region) lie in the neighborhood of the entangled directions. Toward this, we add an angular margin loss [9] to follow latent trajectory in the vicinity of the computed direction $\mathbf{u}$. Formally, let the reference vector be denoted as $\mathbf{z}$ and vectors to be optimized be denoted as $\mathbf{z}_i$ $\forall i\in\{1,...n\}$. We then compute $\mathbf{v_i}=\frac{\mathbf{z}_i-\mathbf{z}}{|\mathbf{z}_i-\mathbf{z}|}$ and compute the following loss:

$$\mathcal{L}_{\text{angle}}=\frac{1}{n}\sum_{i=1}^{n}\arccos\left(\mathbf{v_i}\cdot\mathbf{u}\right)-\theta^{'} \tag{4}$$

where $\theta^{'}$ is the offset parameter. This loss encourages the optimization to stay in the vicinity of the linear trajectory along the entangled direction. Overall we combine Eq. 4 with latent optimization of Eq. 3 to achieve directional latent-optimization. Throughout this paper, we consider latent-optimization without direction as a baseline and refer to it as *Baseline LO*, while referring the presented method as *LO + direction*.

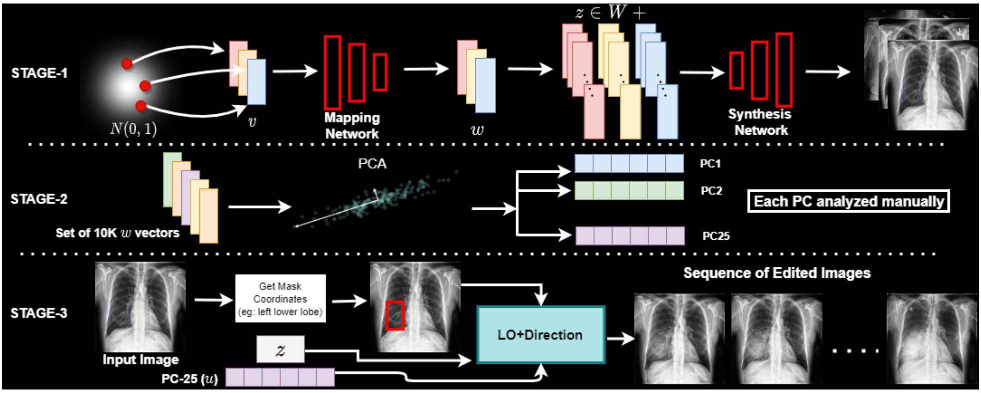


Figure 2: Schematic of the proposed method pipeline demonstrated into three stages. STAGE 1: Training and sampling from StyleGAN, STAGE 2: Manual analysis of PCs in $\mathcal{W}$ space, and STAGE 3: Computing coordinates (manually or with the help of an off-the-shelf lung segmentor) and using the proposed method for image editing.

2.2 Metric for Assessing Disease Progression in Generated Images

Within computer vision literature, metrics like Frechet Inception Distance (FID) [11] are commonly used to assess the overall quality of generated images by a generative model. These metrics, however, are not suitable for evaluating minor, localized changes related to disease progression in the images, as in our framework. In this work, we are motivated by clinical metrics used for disease quantification, such as nonlesion lung volume (NLLV) (lung volume - lesion volume) [7] and Computed Tomography (CT) severity score [27] which fundamentally consider the ratio of abnormal to normal parts of the lung in order to provide a reliable estimate of disease severity. While these clinical metrics are traditionally defined for a single image, we extend them to a multi-image setting: specifically, we present a Pixel-Variation(PV) metric to quantify the progression of specific diseases in user-defined region of interest in an ordered sequence of generated images. We incorporate the following two objectives into the design of the PV metric:

1. **Smoothness**: The sequence of images generated from the optimized latent space should vary to a small degree.

2. **Monotonicity**: The changes taking place in the sequence of images should be in only one direction *i.e.,* either increase or decrease in pixel values.

For a sequence of generated images $\mathbf{x}_i$ $\forall i \in \{1,..n\}$, we first calculate two quantities $d^i_{\text{cmap}} = \mathbf{x}_i$ - $\mathbf{x}_{i-1}$ and $d^i_{\text{rmap}} = \mathbf{x}_i$ - $\mathbf{x}_{\text{ref}}$, respectively denoting consecutive difference map and reference difference map,where $\mathbf{x}_{\text{ref}} = \mathbf{x}_1$. Using these two pixel maps, we then separately calculate the average of positive and negative pixel values resulting in four scalar values: d^i_{c+}, d^i_{c-}, d^i_{r+}, and d^i_{r-}. Since we want change to happen only in the localized region, any changes happening outside the mask are treated negatively. Using these four scalar values, we then measure smoothness S and monotonicity M as:

$$S = \frac{1}{n}\sum_{i=1}^{n} |d^i_{c+}| + |d^i_{c-}| \quad \text{and} \quad M = \frac{1}{n}\sum_{i=1}^{n} d^i_{c+} + d^i_{c-} + d^i_{r+} + d^i_{r-}$$

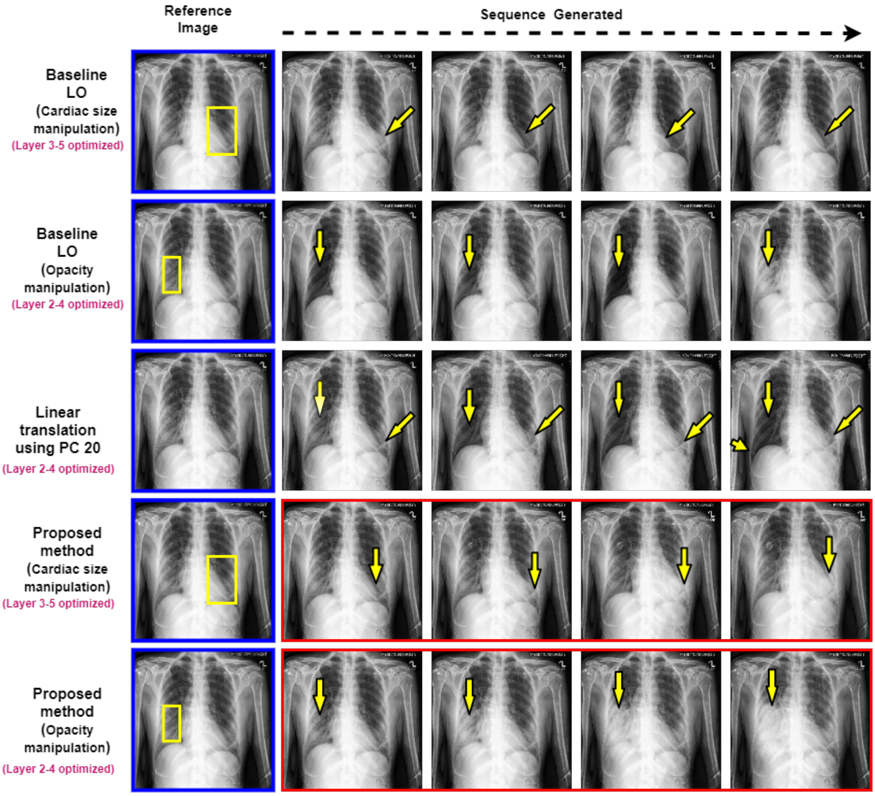
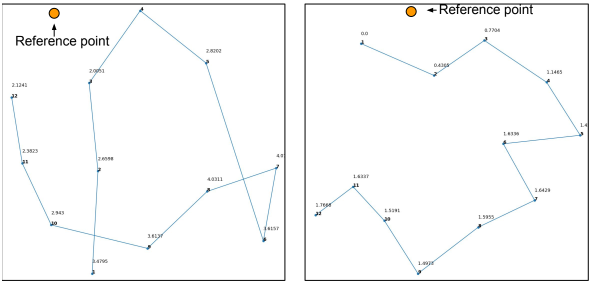

Reference Image
Sequence Generated
Baseline LO (Cardiac size manipulation)
(Layer 3-5 optimized)
Baseline LO (Opacity manipulation)
(Layer 2-4 optimized)
Linear translation using PC 20
(Layer 2-4 optimized)
Proposed method (Cardiac size manipulation)
(Layer 3-5 optimized)
Proposed method (Opacity manipulation)
(Layer 2-4 optimized)

Figure 3: Comparison of LO + direction in the last two rows against Baseline LO in the first two rows. The direction (obtained via PCA analysis) used for the presented framework is demonstrated in the middle row using linear translation. Row 3 corresponds to [12].

We want S to be as small as possible as it accounts for the absolute change in the consecutive images in the trajectory. We want M to be as large as possible as it captures the difference in the desired positive direction against both consecutive image and reference image. As such, our proposed metric PV is defined as the ratio between these two terms, *i.e.,* $PV = \frac{M}{S}$.

3 Experiments

We consider publicly available CheXpert dataset [13] for this study. This dataset contains 224,316 chest radiographs of 65,240 patients. With this dataset, we train state-of-the-art StyleGAN model[22] using all the samples except lateral images. We then use the presented latent-optimization framework to semantically edit X-ray samples with two super-class disease categories of cardiomegaly (enlarged cardiac size) and lung opacity (*e.g.*, consolidation, pneumonia, etc.)[2]. Implementation details can be found in the supplementary material (Section 1), and Fig. 2.1 demonstrates the overview of the proposed pipeline.



Baseline LO

LO + direction

Figure 4: UMAP of latent trajectory for lung opacity for Baseline LO (left) and the presented LO + direction (right).

3.1 Image-Editing Quality

In this section, we provide image quality assessment using expert clinical evaluation and quantitative evaluation using the proposed PV metric. We begin with the qualitative results.

3.1.1 Qualitative Analysis

We demonstrate the capacity of our presented latent optimization framework in Fig. 3. In the first column, we present the reference image. In the rest of the columns, we demonstrate the sequence of generated samples after semantic editing in the pre-defined localized region (green bounding boxes). Unlike most of the work often seen in medical imaging literature [3, 6, 14, 15, 18, 24], the generated samples in this work are of high-quality with clear chest and shoulder anatomy. We present results from the Baseline LO for cardiac size and lung opacity in the first two rows. Although variations are captured in the localized region, these variations are not controlled. As such, we observe irregular variations (*e.g.,* random increase or decrease in the heart size in the first row). In comparison, with the presented LO + direction (in the last two rows), we see a smooth variation for both cardiac size and lung opacity. This allows us to use these edited samples directly as the augmented samples for the respective disease category without manual labeling simply by choosing the last points of the sequence as the monotonicity is maintained. In the third row, we show images generated via linear translation along the directions used to guide the latent optimization As shown, this direction is not perfectly disentangled with variations in both the left and right lobe. Yet, with LO + direction, we can use it to guide our latent-optimization for the respective localized region in each lobes. For other qualitative results, please refer to the provided supplementary material (Section 2).

In Fig. 4, we visualize the trajectory of the latent vectors using UMAP [16] where, compared to the zig-zag path taken by the Baseline LO, the presented LO + direction maintains a coherent trajectory in the latent space.

In our experiments, we also observe that the naive application of Baseline-LO may result in random and objectionable variations irrelevant to the disease of interest. We present one such case in Fig. 5 where cardiac silhouette manipulation leads to collapse of the lung lobe, and the image quality crumbles using Baseline LO. The presented LO + direction maintains the quality of images and monotonous increase in the cardiac size for the same sample.

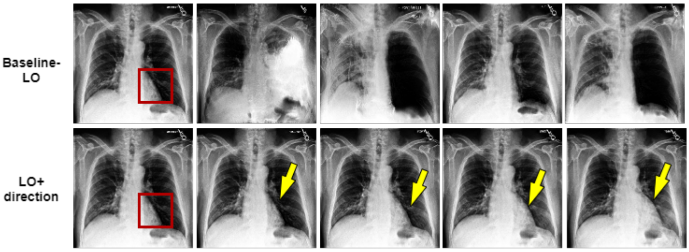


Figure 5: Comparison of the irrelevant variations resulted with Baseline LO (first row) against plausible result from LO + direction (bottom row).

Table 1: Clinical evaluation: First and second column shows the percentage of correctly classified images for each disease category. The third column shows the mean and standard deviation of the overall quality rating (between 1-5) of all images.

	Cardiomegaly	Opacity	Quality rating
Pulmonologist's response	92.00 %	96.00 %	3.60 ± 0.67

3.1.2 Clinical Evaluation

For the clinical evaluation, we selected 25 random images from two groups based on the region of interest. The first group represents the bottom region on the left lobe for obtaining cardiomegaly samples, and the second group represents different regions around the X-ray for obtaining opacity samples. In total, we presented these 50 samples to a board-certified pulmonologist to review the images for features appropriate to two super-class disease categories, namely cardiomegaly and lung opacity, classify them in either or both of the classes, and rate the overall quality of the X-ray on a scale of 1-5 (the higher the better). We provide a sample questionnaire in the supplementary material (Section 3).

We present the overall clinical assessment in Table 1, where we can see that the edited samples resulted from our framework has a meaningful clinical interpretation. In Fig. 6, we present several examples presented to the pulmonologist along with their ratings. As we can see, the synthetic images show similar characteristics to real images, showing enlarged cardiac silhouette in the case of cardiomegaly and opacity in the right lower lobe. This posits that the synthetic images have well-grounded localized changes and can be reliably used for downstream tasks.

3.1.3 Quantitative Results

Table 2 presents the obtained results from the presented PV metric for the Baseline LO compared to the presented LO + direction. Note that a higher value of PV metric means the generated samples are of the desired quality and monotonicity. Each value represents the mean of the PV metric obtained from 50 different normal images subjected to the latent optimization to generate 12 semantically edited images. Since each image's disease region could be located differently, we manually define each image's bounding box. We analyze the metric for 3 different spring-length values (σ) and found that a lower σ value is better suited for the Chest X-ray dataset. For σ values less than 1.0, we get better samples with the presented LO + directions, compared to the Baseline LO for both cardio (representing cardiac size) and lung opacity disease type.

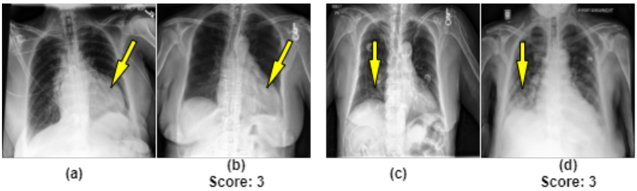


Figure 6: Example of synthesized samples (S) presented to the board-certified pulmonologist along with the real samples (R) from the CheXpert dataset. (a) Cardiomegaly-R;(b) Cardiomegaly-S;(c) Opacity-R;(d) Opacity-S. Markings in the figure are for demonstration purposes only.

Table 2: Quantification of variations in the generated images using PV metric.

Spring-length	Model-type	Disease-type	
		Cardio	Opacity
$\sigma = 0.3$	Baseline LO	0.017 ± 0.37	0.019 ± 0.27
	LO+directions	**0.51** ± 0.20	**0.44** ± 0.14
$\sigma = 1.0$	Baseline LO	-0.05 ± 0.41	-0.03 ± 0.34
	LO+directions	**0.32** ± 0.24	**0.29** ± 0.16
$\sigma = 2.0$	Baseline LO	-0.12 ± 0.14	-0.15 ± 0.19
	LO+directions	**0.11** ± 0.22	**0.14** ± 0.19

3.2 Downstream Application: Chest X-ray Classification

To understand the effect of the synthesized sample, we consider the downstream application of chest X-ray classification. Since we primarily focused our analyses on two super-class disease categories of cardiomegaly and lung opacity, we create an experimental setup first for three-way classification between normal X-ray (N), cardiomegaly (C), and opacity (O). We consider 10K real samples for the normal image and 1K and 5K real samples for the disease category. To investigate the benefit of generated samples, we augment 5K semantically-edited samples (via LO + direction) to the real samples. We compare the presented approach with two different settings: using only real samples and using commonly used random augmentation techniques (rotation, shearing, Gaussian blur, horizontal and vertical flip) to create another augmented set of 5K images. The classification results as presented in the second column of Table 3 clearly demonstrate the benefits of augmented samples. Furthermore, we combine the effect of samples synthesized from our framework and the random augmentation (2.5K samples from each category) and observe the performance to be better to using random augmentation samples alone. This may suggest these two approaches can be combined for further improvement in the generalization performance. We further consider two-way classification between N and C and between N and O. The results are presented in the third and fourth column of Table 3. As before, the benefits of augmenting the synthesized samples can be observed. Note that each value in Table 3 represents mean and standard deviation over four runs with random-seed selection. We consider the same validation and test set for each experimental setup (or each column).

Table 3: Mean AUROC for three-label classification (normal, cardiomegaly, and opacity) in the first column, and AUROC for binary classification in the second (normal *vs.* cardiomegaly) and third (normal *vs.* opacity) column. 5K Aug. refers to the 5K synthesized samples augmented while training the classifier. Values in bold are statistically significant with 95% confidence w.r.t values for random augmentation.

10K Normal + (↓)	N *vs.* C *vs.* O	N *vs.* C	N *vs.* O
1K Disease	75.64 ± 0.25	88.49 ± 0.22	87.73 ± 0.43
1K Disease + 5K Aug.(rand)	75.78 ± 0.29	88.43 ± 0.38	87.90 ± 0.32
1K Disease **+ 5K Aug. (ours)**	**76.60** ± 0.24	**89.01** ± 0.21	**88.60** ± 0.39
1K Disease+5K Aug.(ours+rand)	76.01 ± 0.22	88.83 ± 0.24	88.10 ± 0.34
5K Disease	77.28 ± 0.24	90.01 ± 0.15	90.92 ± 0.21
5K Disease + 5K Aug.(rand)	77.24 ± 0.31	90.10 ± 0.25	91.05 ± 0.25
5K Disease **+ 5K Aug. (ours)**	**77.96** ± 0.22	**91.13** ± 0.12	**91.80** ± 0.16
5K Disease +5K Aug.(ours+rand)	77.52 ± 0.21	90.64 ± 0.29	91.44 ± 0.35

Table 4: Mean AUROC for different values of σ. Values in red are statistically insignificant with 95% confidence w.r.t their corresponding random augmentation values in Table 3

	10K Normal + (↓)	N *vs.* C *vs.* O	N *vs.* C	N *vs.* O
$\sigma = 0.3$	1K Disease **+ 5K Aug.(ours)**	**76.60** ± 0.24	**89.01** ± 0.21	**88.60** ± 0.39
	5K Disease **+ 5K Aug.(ours)**	**77.96** ± 0.22	**91.13** ± 0.12	**91.80** ± 0.16
$\sigma = 1.0$	1K Disease + 5K Aug.(ours)	76.41 ± 0.27	88.90 ± 0.24	88.31 ± 0.23
	5K Disease + 5K Aug.(ours)	77.80 ± 0.21	91.01 ± 0.28	91.51 ± 0.19
$\sigma = 2.0$	1K Disease + 5K Aug.(ours)	74.20 ± 0.34	87.1 ± 0.41	85.90 ± 0.23
	5K Disease + 5K Aug.(ours)	75.11 ± 0.21	88.61 ± 0.23	89.30 ± 0.28

3.3 Ablation Studies

In our experiments, we found σ to be the most important hyper-parameter, therefore we present the ablation study on the effect of different values of σ for different classification cases in Table 4. We observe that although $\sigma = 1.0$ performs better than random augmentations, it performs lower than $\sigma = 0.3$ and is statistically insignificant in some cases . Ablation of the σ for quantifying variations in the generated images using PV metric is provided in Table 2. More details on ablation of σ and similar PV metric analysis for $\theta^{'}, \alpha, \beta, \gamma$ along with their trajectory UMAPs and images for a sample case is presented in Section 5 of the supplementary material. For all the analyses, we obtain the best results with $\sigma = 0.3$ and $\theta^{'} = 35°$.

4 Conclusion

In this paper, we present a semantic image editing framework via latent-optimization with explicit semantic directions for medical image augmentation producing disease-aware variations in the user-defined region of interest. We demonstrate the benefits of the presented framework via the augmented samples, which we analyzed qualitatively, clinically and quantitatively using the presented novel PV metric, by improving the performance of a downstream disease classification task. Finally, finding a suitable disentangled direction for semantic editing is an active and underexplored area in medical imaging. We believe this framework also provides a stride in this direction. One interesting future research avenue is to consider a smooth variation of the augmented samples generated by our framework to build models for studying disease progression.

While the proposed method demonstrated some benefits in medical imaging, it does comes with some limitations. First, It assumes that StyleGAN can generate a diversity of images. If the StyleGAN is trained with truncation trick to favor image quality over diversity, then the proposed method might face difficulty in discovering latent points that may generate images with required changes in the masked region. Second the latent optimization has a high run time, taking up to a minute till convergence, a significant hurdle for large-scale adoption. Third, the performance of the proposed pipeline depends on whether the latent point is placed in a well-behaved region of the latent space. The third limitation could be more pressing as it could lead to the discovery of suboptimal latent points, resulting in unwanted artifacts in the generated images. One such example can be observed in Fig. 3 (row 4), which has a bright circular region on the upper left lobe as artifacts. In some cases, such ill-positioned latent points may hamper the overall optimization as well such as negligible changes in the masked region. We provide one such example in Fig. 11 of the supplementary material. Overall, addressing these challenges could be an interesting future avenue.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8296–8305, 2020.

[2] J. Irvin et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. CoRR, abs/1901.07031, 2019. URL http://arxiv.org/abs/1901.07031. _eprint: 1901.07031.

[3] Saleh Albahli. Efficient GAN-based Chest Radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia. Int J Med Sci, 17(10):1439–1448, June 2020. ISSN 1449-1907. doi: 10.7150/ijms.46684. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7330663/.

[4] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In 6th International Conference on Learning Representations, ICLR 2018, 2018.

[5] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating Highly Realistic Images of Skin Lesions with GANs. arXiv:1809.01410 [cs, eess], September 2018. URL http://arxiv.org/abs/1809.01410. arXiv: 1809.01410.

[6] Debangshu Bhattacharya, Subhashis Banerjee, Shubham Bhattacharya, B. Uma Shankar, and Sushmita Mitra. GAN-Based Novel Approach for Data Augmentation with Improved Disease Classification. In Om Prakash Verma, Sudipta Roy, Subhash Chandra Pandey, and Mamta Mittal, editors, Advancement of Machine Intelligence in Interactive Medical Image Analysis, Algorithms for Intelligent Systems, pages 229–239. Springer, Singapore, 2020. ISBN 9789811511004. doi: 10.1007/978-981-15-1100-4_11. URL https://doi.org/10.1007/978-981-15-1100-4_11.

[7] Wenli Cai, Tianyu Liu, Xing Xue, Guibo Luo, Xiaoli Wang, Yihong Shen, Qiang Fang, Jifang Sheng, Feng Chen, and Tingbo Liang. CT Quantification and Machine-learning Models for Assessment of Disease Severity and Prognosis of COVID-19 Patients. Acad Radiol, 27(12):1665–1678, December 2020. ISSN 1878-4046. doi: 10.1016/j.acra.2020.09.004.

[8] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic Latent Space Interpolation for Unpaired Image-To-Image Translation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 2403–2411, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00251. URL https://ieeexplore.ieee.org/document/8954444/.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4690–4699, 2019.

[10] Prashnna Kumar Gyawali, Zhiyuan Li, Sandesh Ghimire, and Linwei Wang. Semi-supervised learning by disentangling and self-ensembling over stochastic latent space.

In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 766–774. Springer, 2019.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs, stat], January 2018. URL http://arxiv.org/abs/1706.08500. arXiv: 1706.08500.

[12] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. arXiv:2004.02546 [cs], December 2020. URL http://arxiv.org/abs/2004.02546. arXiv: 2004.02546.

[13] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 590–597, 2019.

[14] Zhaohui Liang, Jimmy Xiangji Huang, Jun Li, and Stephen Chan. Enhancing Automated COVID-19 Chest X-ray Diagnosis by Image-to-Image GAN Translation. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 1068–1071, December 2020. doi: 10.1109/BIBM49941.2020.9313466.

[15] Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In Medical Imaging 2018: Image Processing, volume 10574, page 105741M. International Society for Optics and Photonics, 2018.

[16] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.

[17] Aakash Saboo, Sai Niranjan, Kai Dierkes, and Hacer Keles. Towards disease-aware image editing of chest x-rays. In 4th Workshop on Medical Imaging meets Neurips at NeurIPS 2020, 2020.

[18] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks. arXiv:1712.01636 [cs], February 2018. URL http://arxiv.org/abs/1712.01636. arXiv: 1712.01636.

[19] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. Scientific reports, 9(1):1–9, 2019.

[20] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9243–9252, 2020.

[21] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine Andriole, and Mark Michalski.

Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks. arXiv:1807.10225 [cs, stat], September 2018. URL http://arxiv.org/abs/1807.10225. arXiv: 1807.10225.

[22] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. CoRR, abs/1812.04948, 2018. URL http://arxiv.org/abs/1812.04948. _eprint: 1812.04948.

[23] Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pages 1251–1255. IEEE, 2019.

[24] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Placido Rogerio Pinheiro. CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. IEEE Access, 8:91916–91923, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2994762. URL http://arxiv.org/abs/2103.05094. arXiv: 2103.05094.

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[26] Mengyu Yang, David Rokeby, and Xavier Snelgrove. Mask-guided discovery of semantic manifolds in generative models. In 4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020, 2020.

[27] Ran Yang, Xiang Li, Huan Liu, Yanling Zhen, Xianxiang Zhang, Qiuxia Xiong, Yong Luo, Cailiang Gao, and Wenbing Zeng. Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. Radiology: Cardiothoracic Imaging, 2(2):e200047, April 2020. doi: 10.1148/ryct.2020200047. URL https://pubs.rsna.org/doi/full/10.1148/ryct.2020200047. Publisher: Radiological Society of North America.