# Analysing the effectiveness of a generative model for semi-supervised medical image segmentation

**Margherita Rosnati** MARGHERITA.ROSNATI12@IMPERIAL.AC.UK
*BioMedIA Group, Department of Computing, Imperial College London, UK*

**Fabio De Sousa Ribeiro** F.DE-SOUSA-RIBEIRO@IMPERIAL.AC.UK
*BioMedIA Group, Department of Computing, Imperial College London, UK*

**Miguel Monteiro** MIGUEL.MONTEIRO@IMPERIAL.AC.UK
*BioMedIA Group, Department of Computing, Imperial College London, UK*

**Daniel Coelho de Castro** DACOELH@MICROSOFT.COM
*Microsoft Research, UK*
*BioMedIA Group, Department of Computing, Imperial College London, UK*

**Ben Glocker** B.GLOCKER@IMPERIAL.AC.UK
*BioMedIA Group, Department of Computing, Imperial College London, UK*

## Abstract

Image segmentation is important in medical imaging, providing valuable, quantitative information for clinical decision-making in diagnosis, therapy, and intervention. The state-of-the-art in automated segmentation remains supervised learning, employing discriminative models such as U-Net. However, training these models requires access to large amounts of manually labelled data which is often difficult to obtain in real medical applications. In such settings, semi-supervised learning (SSL) attempts to leverage the abundance of unlabelled data to obtain more robust and reliable models. Recently, generative models have been proposed for semantic segmentation, as they make an attractive choice for SSL. Their ability to capture the joint distribution over input images and output label maps provides a natural way to incorporate information from unlabelled images. This paper analyses whether deep generative models such as the SemanticGAN are truly viable alternatives to tackle challenging medical image segmentation problems. To that end, we thoroughly evaluate the segmentation performance, robustness, and potential subgroup disparities of discriminative and generative segmentation methods when applied to large-scale, publicly available chest X-ray datasets.

## 1. Introduction

Deep learning has shown promising results in medical image segmentation (Ronneberger et al., 2015). Applications include quantifying disease progression (Liu et al., 2020), lesion volumes (Robben et al., 2020), tumour progression (Abdelazeem et al., 2020) and radiotherapy planning (Oktay et al., 2020). However, large annotated datasets for training are scarce: despite data being routinely collected for all patients, expert annotations are prohibitively expensive and time-consuming to obtain in practice. As a result, models are often trained on smaller datasets. Consequently, models may not generalise well, with an often observed drop in performance when the data characteristics change—known as domain shift (Quinonero-

Candela et al., 2008). In addition, potential biases in training data could cause deep learning methods to exacerbate health disparities (Obermeyer et al., 2019; Adamson and Smith, 2018).

Semi-supervised learning (SSL) (Chapelle et al., 2006) seeks to solve the problem of labelled data scarcity by leveraging information from unlabelled data in addition to the labelled data. Recent literature on SSL for segmentation focused on discriminative methods, mainly by augmenting the labelled training data through pseudo-labelling (Ke et al., 2020) or constraining models' latent representation by enforcing consistent predictions within similar datapoints (Ouali et al., 2020). Methods based on these concepts have been extensively studied in the medical imaging context (Shurrab and Duwairi, 2022; Jiao et al., 2022).

Advances in SSL also span generative methods (Sajun and Zualkernan, 2022). The works by Wei et al. (2018) and Li et al. (2017) learn the distribution of the input data $\mathcal{X}$ to train a classifier more robustly. Recently, Li et al. (2021) expanded the modelling of $p(x)$ to the joint distribution $p(x, y)$ of images $x$ and their semantic segmentations $y$ for SSL. Their method showed promising preliminary results on generalisations to new tasks and datasets, characteristics particularly desirable in medical imaging.

This work examines the aptitude of SSL deep generative models that learn the joint distribution of $p(x, y)$ to the challenging real-world medical image segmentation context. We conduct an extensive empirical analysis on SemanticGAN (Li et al., 2021) as a representative example, compared to a state-of-the-art fully supervised network and a semi-supervised discriminative ablation study on SemanticGAN for chest X-ray lung segmentation, to substantiate the following:

- We provide thorough experimental evidence for SemanticGAN's generalisability for in- and out-of-domain tasks according to segmentation metrics and downstream disease classification performance;
- We uncover that the model's robustness cannot be explained by adversarial training alone;
- We characterise the model's strengths and weaknesses regarding subgroup disparities and biases in the training data.

## 2. Related work

**SSL for semantic segmentation** Recent advances in SSL for semantic segmentation can be broadly categorised into two sometimes overlapping clusters: pseudo-label based and consistency constraint based. The former cluster use supervised learning architectures trained multiple times over pseudo-labels, which are labels produced by model predictions (Olsson et al., 2021; Chen et al., 2021; Wang et al., 2022). A limitation of these models is that they do not leverage the unlabelled datapoints for which their predictions are not confident. The latter uses auxiliary tasks to generate a separable and information-dense latent representation of the data where similar datapoints are close (Ouali et al., 2020; Liu et al., 2022). For example, an idea borrowed from generative adversarial networks (GAN) is to use a discriminator to constrain the segmentation model to produce consistent segmentations (Hung et al., 2018). These methods all suffer from poor generalisation to unseen data, and sometimes underperform fully supervised methods (Singh et al., 2008; Yang et al., 2021).

**Generative methods in SSL** A stream of generative methods for SSL uses variational autoencoders (VAEs), modelling $p(x, y)$ through various latent variable models (Kingma et al., 2014; Ehsan Abbasnejad

et al., 2017; Joy et al., 2020). However, VAEs are rarely used in per-pixel classification.

In contrast, there are several GAN-inspired methods for generative SSL. Donahue et al. (2016) devised the "bi-directional GAN" method, where an encoder maps the image distribution to the GAN latent space. Further works (Dumoulin et al., 2016; Kumar et al., 2017) use the latent space as the classification feature space. These works generate a compact representation of the data distribution of the images $p(x)$ and use it to predict the image label in a supervised manner. Other authors employed both generative modelling and adversarial training to learn the image label. Using architectures derived from TripleGAN (Li et al., 2017), Dong and Lin (2019) and Wu et al. (2019) propose to learn $p(x)$ through a GAN generator. Samples from the GAN are then concatenated to labels discriminatively generated by a classifier and fed to a discriminator for adversarial training. To the best of our knowledge, no GAN-based method other than Li et al. (2021) learns the joint distribution $p(x, y)$. For this reason, our work focuses on the latter.

**SSL in medical imaging** The literature on SSL methods in medical imaging broadly follows that of semantic segmentation, with pseudo-labelling methods (Han et al., 2022), consistency regularisation methods (Basak et al., 2022) and adversarial learning (Peiris et al., 2021). An in-depth review can be found at Jiao et al. (2022). Generative methods are seldom used. Liu et al. (2019) used a GAN to reconstruct images as an auxiliary task for diabetic retinopathy screening, and Zhang et al. (2022) used an architecture similar to TripleGAN to detect Parkinson's Disease. To our knowledge, no previous work aims to assess methods that model the joint distribution of images and labels in SSL.

## 3. Methods

The SemanticGAN Li et al. (2021) authors hypothesised that modelling the joint distribution $p(x, y)$ of images and segmentations yields superior robustness to domain changes in comparison to discriminative methods, or methods only modelling $p(x)$. We aim to verify this hypothesis in a real-world medical imaging context. Specifically, we compare a generative SSL method, SemanticGAN (Li et al., 2021), to a state-of-the-art fully supervised semantic segmentation architecture, DeepLabV3 (Chen et al., 2017) and to an ablation study on SemanticGAN. Our proposed ablation study pertains to the generative arm of SemanticGAN to determine whether adversarial training is sufficient for SemanticGAN's success.

### 3.1. SemanticGAN

SemanticGAN bases its architecture on StyleGAN2 (Karras et al., 2020). Traditional GANs sample a noise vector $z \sim \mathcal{N}(0, I)$ and transform it through an upsampling architecture into an image. In contrast, StyleGAN2 transforms the noise vector into a higher dimensional style vector $w \in \mathcal{W}$ that is fed to the upsampling architecture at different stages, similarly to style transfer architectures. StyleGAN2 also borrows ideas from ResNet as it uses skip-connections to generate images of increasing resolution. SemanticGAN modifies StyleGAN2 by using a second set of skip-connections to generate the image semantic segmentation, and adding a multi-scale patch-based discriminator (Wang et al., 2018) for the segmentation. The method also adds an encoder from the image space into the $\mathcal{W}$ space (Richardson et al., 2021).

The model is trained against two discriminator models: the first, $D_r$, determines if the image is generated or real, and the second, $D_m$, determines if an image and segmenta-

tion pair is generated or real. Following the original notation (Li et al., 2021), the objectives are defined as:

$$\mathcal{L}_G = \mathop{\mathbb{E}}_{\substack{z\sim p(z)\\(\tilde{x},\tilde{y})=G(z)}} [\log(1 - D_r(\tilde{x})) + \log(1 - D_m(\tilde{x},\tilde{y}))], \quad (1)$$

$$\mathcal{L}_{D_r} = - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_u} [\log D_r(x)] - \mathop{\mathbb{E}}_{\substack{z\sim p(z)\\\tilde{x}=G_X(z)}} [\log(1 - D_r(\tilde{x}))], \quad (2)$$

$$\mathcal{L}_{D_m} = - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_l} [\log D_m(x,y)] - \mathop{\mathbb{E}}_{\substack{z\sim p(z)\\(\tilde{x},\tilde{y})=G(z)}} [\log(1 - D_m(\tilde{x},\tilde{y}))], \quad (3)$$

where $G$ is the generator, $\mathcal{D}_u$ and $\mathcal{D}_l$ are the unlabelled and labelled datasets, respectively, and $(x, y)$ and $(\tilde{x}, \tilde{y})$ are resp. real and generated samples.

Lastly, the model is also composed of an encoder $E$, which maps images to latent style representations compatible with StyleGAN2. The latter is trained with both an image reconstruction loss $\mathcal{L}_u$ and a semantic segmentation reconstruction loss $\mathcal{L}_s$:

$$\mathcal{L}_u = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\mathcal{L}_{\text{LPIPS}}(x, G_X(E(x))) + \lambda_1 \|x - G_X(E(x))\|_2^2], \quad (4)$$

$$\mathcal{L}_s = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_l} [\mathbf{H}(y, G_Y(E(x))) + \mathbf{DC}(y, G_Y(E(x)))], \quad (5)$$

where $\mathcal{L}_{\text{LPIPS}}(\cdot,\cdot)$ is the *Learned Perceptual Image Patch Similarity* distance (Zhang et al., 2018), $\mathbf{H}$ is the cross entropy loss, $\mathbf{DC}$ is the Dice coefficient loss (Isensee et al., 2018) and $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$. The model is trained in two steps. Firstly, the image and segmentation generative branch is trained along with the discriminators. Secondly, the weights of the generative branch are frozen, and the encoder is trained. We use the architectures in the original paper, where the modified StyleGAN2 generator has 31M parameters, the StyleGAN2 discriminator $D_r$ has 28M, the multi-scale patch-based discriminator $D_m$ has 8M, and the encoder $E$ has 7.4B parameters.

### 3.2. Fully supervised baseline

We compare SemanticGAN with DeepLabV3 on a 101-layer ResNet (He et al., 2016) backbone, totalling 59M parameters. DeepLabV3 is considered state-of-the-art in semantic segmentation; its success based on augmenting the field of view of the network through atrous convolutions. The architecture was optimised with a similar loss function to the SemanticGAN encoder $\mathcal{L}_s$ (Eq. (5)) for comparability:

$$\mathcal{L}_{\text{SupOnly}} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_l} [\mathbf{H}(y, \text{DL}(x)) + \mathbf{DC}(y, \text{DL}(x))], \quad (6)$$

where DL stands for DeepLabV3. We call this supervised baseline SupOnly.

### 3.3. Non-Generative Adversarial Network

For the ablation study, we employ the architecture of DeepLabV3 to produce segmentation and train the model adversarially using SemanticGAN's architecture of $D_m$, which together have 67M parameters. During training, we pass labelled or unlabelled images through the CNN architecture. We concatenate the resulting segmentation with the image and pass it through a discriminative model, charged with determining whether the pair comes from the labelled data distribution. In mathematical terms, the objective functions are (cf. Eqs. (1) and (3)):

$$\mathcal{L}_{\text{Seg}} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\log(1 - D_m(x, \text{DL}(x)))], \quad (7)$$

$$\mathcal{L}_{\text{Adv}} = - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_l} [\log D_m(x,y)] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}} [\log(1 - D_m(x, \text{DL}(x)))], \quad (8)$$

where $\mathcal{L}_{\text{Seg}}$ optimises the weights of the discriminative segmentation model, and $\mathcal{L}_{\text{Adv}}$ optimises the weights of the adversarial model. We call this adversarially trained method SemanticAN.

By design, SemanticAN is an ablation on SemanticGAN, testing whether the adversarial training is sufficient for an SSL model to generalise well, or whether learning the distribution of $p(x)$ is beneficial.

In Appendix B, we also compare SemanticGAN to SupOnly and SemanticAN based on a U-Net backbone (Ronneberger et al., 2015) (578k param.) to evaluate the requirement of large models.

## 4. Datasets

In order to compare and contrast the applicability of different approaches, we test them on the task of chest X-ray lung segmentation. Chest X-rays are one of the most common radiological examinations, and automatically extracted features from anatomical regions such as the lungs can support clinical decisions.

We use the ChestX-ray8 (Wang et al., 2017) (n=108k) and CheXpert (Irvin et al., 2019) (n=76k) unlabelled datasets and the NIH (Tang et al., 2019) (n=95), JSRT (Van Ginneken et al., 2006) (n=431), and Montgomery (Jaeger et al., 2014) (n=138) labelled datasets, where the NIH dataset is an annotated subset of the ChestX-ray8 dataset. A detailed description of the datasets can be found in Appendix table A.I, while thumbnails of the datasets can be found in Appendix figures D and A.IV, and raining details in Appendix A.

## 5. Experiments

### 5.1. Model performance and in- and out-of-domain generalisation

We assess the resilience of generative segmentation networks to dataset changes by training SemanticGAN with the ChestX-ray8 unlabelled dataset and the JSRT labelled dataset, and testing the performance on the hold-out set of JSRT, the NIH dataset, and the Montgomery dataset. In this manner, we test the model on in-domain data for the labelled dataset, in-domain data for the unlabelled dataset, and out-of-domain data. We calculate the Dice coefficient, precision, recall, average surface distance, and Hausdorff distance in pixels for each population. The precision and recall highlight methods prone to over- or under-segmentation, whereas the average surface distance is sensitive to the distance between segmentations, and the Hausdorff distance is sensitive to far away outliers. We compare the performance of SemanticGAN to that of SupOnly and SemanticAN, where SupOnly is only trained on the JSRT labelled dataset.

The results summarised in Table 1 show SupOnly consistently outperformed SemanticGAN and SemanticAN when tested on JSRT and Montgomery. When tested on NIH, SemanticGAN outperformed SupOnly in terms of recall and Hausdorff distance, whereas Dice score, precision and average surface distance were inconclusive when comparing the two. These findings are in line with a stream of literature (Yang et al., 2021) claiming that SSL methods do not consistently outperform fully supervised methods.

Nevertheless, SemanticGAN produced Dice, precision and recall scores on average greater than 90%, and an average surface distance of at most 8.8 pixels. The results indicate that the performance of SemanticGAN on datasets from a different

Table 1: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset.

| | Dice | Precision | Recall | Avg. surface distance | Hausdorff distance |
|---|---|---|---|---|---|
| | JSRT (labelled in-domain) | | | | |
| SupOnly | **99.3 ± 0.4** | **99.1 ± 0.4** | **99.5 ± 0.5** | **0.3 ± 0.2** | **1.1 ± 0.4** |
| SemanticAN | 91.2 ± 5.2 | 93.3 ± 3.7 | 89.8 ± 7.6 | 3.7 ± 2.0 | 14.7 ± 7.1 |
| SemanticGAN | 98.2 ± 1.3 | 98.3 ± 1.5 | 98.1 ± 1.2 | 0.7 ± 0.6 | 2.5 ± 2.4 |
| | NIH (unlabelled in-domain) | | | | |
| SupOnly | **90.3 ± 8.1** | **95.3 ± 4.6** | 86.5 ± 11.4 | **4.8 ± 4.8** | 25.5 ± 22.7 |
| SemanticAN | 72.4 ± 19.6 | 75.4 ± 18.8 | 71.2 ± 21.2 | 10.9 ± 7.13 | 35.5 ± 17.1 |
| SemanticGAN | 90.1 ± 3.4 | 89.9 ± 6.5 | **90.9 ± 5.0** | 4.9 ± 3.3 | **18.2 ± 16.0** |
| | Montgomery (out-of-domain) | | | | |
| SupOnly | **96.5 ± 2.3** | **98.8 ± 0.7** | **94.4 ± 4.0** | **1.3 ± 0.7** | **5.5 ± 7.2** |
| SemanticAN | 57.9 ± 33.4 | 62.4 ± 31.0 | 55.5 ± 35.1 | 19.3 ± 15.1 | 54.8 ± 38.9 |
| SemanticGAN | 91.3 ± 5.6 | 92.6 ± 7.9 | 90.4 ± 4.2 | 8.8 ± 10.5 | 34.2 ± 39.1 |

distribution is consistent, although not consistently superior to baselines.

SemanticAN underperformed SemanticGAN and the supervised baseline, whether presented with scans from the distribution of a labelled training set, unlabelled training set or a new dataset. Its sub-par performance on NIH suggests that the model did not exploit the unlabelled dataset $\mathcal{D}_u$, and overfit to the training data distribution $\mathcal{D}_l$. We find that the adversarial training alone is insufficient to learn the imaging data distribution and distorts the training signal, resulting in a weaker model than both SemanticGAN and the supervised baseline.

### 5.2. Downstream task performance for disease classification

As a secondary indicator of the model performance, we evaluate the segmentation algorithm on a downstream task. This evaluation method benefits from not requiring ground truth segmentations, allowing us to test on larger datasets, such as CheXpert and ChestX-ray8. We choose the disease classification downstream task, where we use the segmentation algorithm to mask the background to the lungs and detect different lung abnormalities.

For this purpose, we compare the performance of a classification model trained on original unmasked images and images masked with SupOnly, Semantic AN, and SemanticGAN from the CheXpert dataset. Unfortunately, we cannot compare these performances to that of the CNN trained on images segmented by clinicians due to the absence of such segmentations in the CheXpert dataset. The implementation details can be found in Appendix A. Within the reported 14 pathologies, we pick three clearly visible within the lung area and with a significant prevalence: pleural effusion (35.6%), lung opacity (47.8%) and oedema (18.9%), where lung opacity and pleural effusion co-occur in 28.5% of patients, lung opacity and

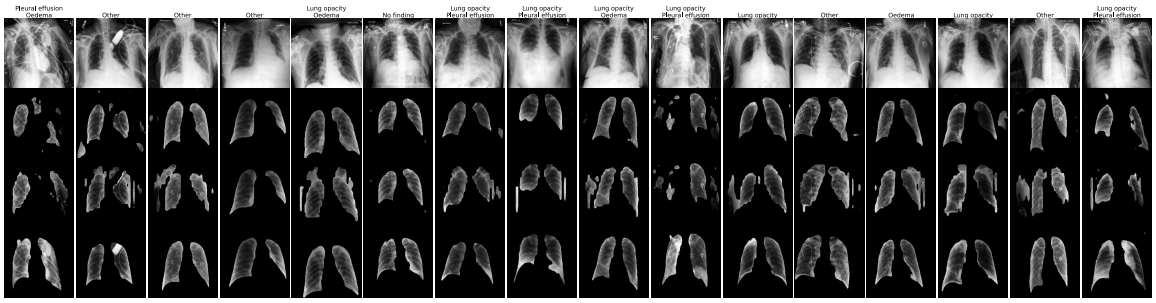



Figure 1: CheXpert images masked by models trained on ChestX-ray8 and JSRT. First row: unmasked image; second row: SupOnly segmentation; third row: SemanticAN segmentation; fourth row: SemanticGAN segmentation. Titles: pathology label.

oedema co-occur in 16.5% of patients, pleural effusion and oedema co-occur in 15.6% of patients, and all three pathologies co-occur in 9.5% of patients. We hypothesise that the information in the lung area is sufficient to classify the image correctly. Hence, a perfectly segmented image should be as predictive as the unmasked image and more predictive than any erroneously segmented one. We train a binary classifier for each pathology and segmentation method and report the area under the receiver operator characteristic curve (AUROC) on the test images segmented with the same method as the training dataset. We provide confidence intervals by training the binary classifier with five different seeds. In addition, we inspect a sample of the segmented images for a qualitative assessment.

Figure 1 shows examples of patients' scans segmented with SemanticGAN and the two baselines. Both SupOnly's and SemanticAN's segmentations have irregular and disconnected shapes. They often mistake some low-intensity background areas for lungs, and under-segment high-intensity areas in the lungs, such as medical devices. SemanticGAN's segmentations are anatomically plausible in shape across patients. However, the segmentation is sometimes misaligned with the lung – for example on the last row of Figure 1, the top-left segmentation starts before the lung area.

Table 2 shows the results of the classification tasks. In contrast with the results in Section 5.1, SemanticGAN outperformed both SupOnly and SemanticAN for the classification of all tasks but Pleural Effusion, where SupOnly and SemanticGAN perform similarly. Paired with the qualitative assessment, we deduce that although SemanticGAN scores low on the segmentation metrics in Section 5.1, its segmentations are consistent in shape and may more robustly capture the anatomical region of interest for disease classification.

In addition, an investigation of how the pathologies' co-occurrence affect predictions found that all models predicted worse for patients with co-occurrences than for patients without co-occurrence. Our intuition is that the more pathologies are present in the image, the noisier the signal for each individual pathology, and the harder it can be for a given model to classify the (masked) images. In addition, the hierarchy of performance between SupOnly, SemanticAN, SemanticGAN and Unmasked shown in Table 2 was observed for patient subgroups with multiple pathologies.

Table 2: Diagnosis classification AUROC for images masked using different segmentation models on the CheXpert dataset (out-of-domain). Reported as mean ± standard deviation over the dataset.

| | Opacity | Effusion | Oedema |
|---|---|---|---|
| SupOnly | 67.4 ± 0.2 | 82.0 ± 0.1 | 77.8 ± 0.2 |
| SemanticAN | 66.6 ± 0.2 | 78.9 ± 0.2 | 77.2 ± 0.1 |
| SemanticGAN | **67.9 ± 0.1** | **82.2 ± 0.2** | **78.1 ± 0.2** |
| Unmasked | **70.1 ± 0.3** | **85.0 ± 0.2** | **79.1 ± 0.1** |

Finally, the superior classification performance of the classifier trained on unmasked images indicates that lung segmentations are imperfect for all methods, or other image regions may play an important role in the prediction of disease.

### 5.3. Performance across subgroups and training bias

**Subgroup performance** Figure 2 shows the stratification by the patient's biological sex of the Dice similarity coefficient derived in Section 5.1. A corresponding numerical table can be found in Appendix Table A.V. The relative model performance for females versus males was mixed for all models, except for SemanticAN. The latter performed statistically significantly better for males than females for all datasets. On the other hand, SupOnly performed comparably for the two subgroups for all datasets, and SemanticGAN performed comparably for the subgroups for all datasets but Montgomery, where it performed better for males. As SemanticAN is overfitting the labelled training data, it is unsurprising that it carries the same bias as on the training group on other groups. SemanticGAN showed to be more robust to biases than SemanticAN, yet worse than the fully supervised baseline.

When comparing models for each subgroup, SemanticAN underperformed both SupOnly and SemanticGAN for all subgroups and datasets, and SupOnly outperformed SemanticGAN for all subgroups but females in NIH, where they performed comparably. These results are consistent with the findings in Section 5.1.

A similar analysis stratifying the results of Section 5.2 can be found in Appendix C.

**Training Bias** A follow-up question to the subgroup stratification experiment is how stronger population biases in the training dataset affect the model performance over protected subgroups (Larrazabal et al., 2020). Specifically, we wish to learn whether the training of different parts of SemanticGAN with biased data affects the model performance; and whether the performance is more sensitive to a bias in either the labelled or the unlabelled datasets.

On that account, we generate a biased unlabelled dataset $\widetilde{\mathcal{D}_u}$ from ChestX-ray8, and a labelled biased dataset $\widetilde{\mathcal{D}_l}$ from JSRT, both composed of only the biological male, removing all biological females from the original datasets. We perform three sets of experiments, recapitulated in Appendix table A.IV. Firstly, we train a model solely on biased data, illustrating the most extreme circumstances. We call it the "full bias" model. Secondly, we train two models using the biased datasets at two different training phases. We train one model by train-

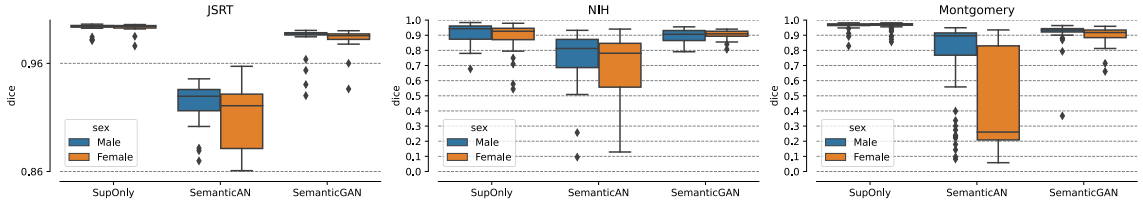



Figure 2: Stratification of segmentation results (Sections 5.1) by biological sex.

ing the generator $G$ with the biased datasets $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}_u} \cup \widetilde{\mathcal{D}_l}$ and then the encoder $E$ with the original datasets $\mathcal{D}$. We train a second model by training $G$ with the original datasets $\mathcal{D}$ and $E$ with the biased datasets $\widetilde{\mathcal{D}}$. We call these the "biased generator" and "biased encoder" models respectively. Thirdly, we train one model with biased labelled data $\widetilde{\mathcal{D}_l}$ and original unlabelled data $\mathcal{D}_u$ for both $G$ and $E$, and a second model with original labelled data $\mathcal{D}_l$ and biased unlabelled data $\widetilde{\mathcal{D}_u}$. We call these the "biased labelled dataset" and "biased unlabelled dataset" models respectively.

For each iteration, we test the model on the same unbiased test datasets, JSRT, NIH and Montgomery, and compare their performance to the model from Sections 5.1 and 5.2, trained on the original dataset $\mathcal{D}$—the "control" model.

Figure 3 presents the differences in Dice scores when training SemanticGAN with biases in different parts of the process. A detailed table with numerical values can be found in Appendix Table A.VI. Overall, the training bias affected the performance of the JSRT dataset more than on the other datasets. For JSRT and biological females, the full bias, biased encoder and labelled data models performed significantly worse than the control. However, biasing generator and unlabelled data did not harm the model performance. In addition, the female bias did not affect the male subgroup performance. However, for the NIH and Montgomery datasets, the discrepancy in model performance was smaller. The control model performed better than the full bias and biased unlabelled dataset models for females, and the biased encoder model for females in NIH.

The in-domain dataset results on JSRT for labelled data and NIH for unlabelled data were sensitive to biases in their training domain and the biased encoder. Intuitively, these results are coherent with the training paradigm, as the encoder explicitly minimises the difference between the datasets and the models' outputs. The sensitivity to the biased unlabelled dataset in the out-of-domain dataset results also substantiates the hypothesis that SemanticGAN's generalisation properties derive from its training on unlabelled data. Consequently, if the model is presented with data significantly different from the unlabelled dataset, we expect SemanticGAN to decrease in performance.

## 6. Conclusion

In conclusion, this work provides a thorough analysis of the segmentation performance, robustness, and potential subgroup disparities of discriminative and generative segmentation methods when applied to large-scale, publicly available chest X-ray datasets.

We found that SemanticGAN generates consistent predictions in shape (Section 5.2) and accuracy (Section 5.1), performing better than baselines on downstream tasks (Section 5.2). Our experiments on SemanticAN

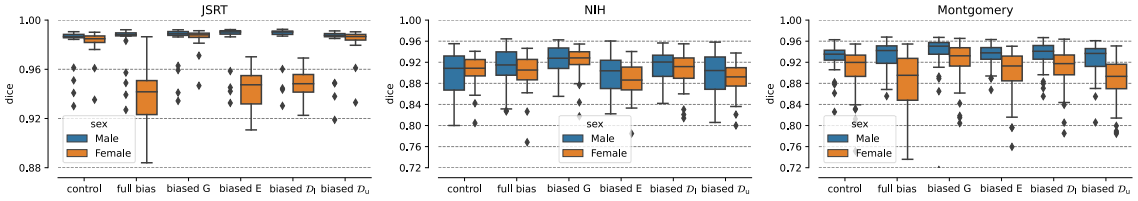



Figure 3: Impact of biological sex bias on model performance.

in Section 5.1, the ablation study model, revealed that SemanticGAN's performance is due to its ability to learn the joint image and segmentation distribution $p(x, y)$.

The experiments on biased training in Section 5.3 highlighted the model's reliance on the unlabelled dataset for generalisation to out-of-domain datasets. In addition, counter-intuitively, the most significant relative drop in performance was observed in the labelled in-domain setting. Further work should aim to understand the model's weaknesses to different biased training data.

Although SemanticGAN showed strong and consistent performance, the fully-supervised baseline consistently scored higher on segmentation metrics. An interesting follow-up study would be to test the hypothesis that the method's strengths outweigh its weaker performance in an active learning regime, whereby only the most informative examples are annotated.

Our findings suggest that generative models are particularly well-suited when shape consistency is a critical desideratum. For other circumstances, the results reported in this work show that the method examined generalises well and does not exacerbate existing biases. However, SemanticGAN sometimes fails to perform better than a baseline that is easier to train and achieves excellent segmentation scores. We conclude that generative approaches to medical image segmentation have potential and should be investigated in future work. They may become a viable alternative to discriminative models, in particular, due to their ability to incorporate information from unlabelled data.

## Acknowledgments

MR is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

## References

Rania M Abdelazeem, Doaa Youssef, Jala El-Azab, Salah Hassab-Elnaby, and Mostafa Agour. Three-dimensional visualization of brain tumor progression based accurate segmentation via comparative holographic projection. *Plos one*, 15(7):e0236835, 2020.

Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154 (11):1247–1248, 2018.

Hritam Basak, Rajarshi Bhattacharya, Rukhshanda Hussain, and Agniv Chatterjee. An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation. *arXiv preprint arXiv:2202.00677*, 2022.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, 2006.

Dali Chen, Dingpeng Sun, Jun Fu, and Shixin Liu. Semi-supervised learning

framework for aluminum alloy metallographic image segmentation. *IEEE Access*, 9:30858–30867, 2021.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Jinhao Dong and Tong Lin. Margingan: Adversarial training in semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

M Ehsan Abbasnejad, Anthony Dick, and Anton van den Hengel. Infinite variational autoencoder for semi-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5888–5897, 2017.

Kai Han, Lu Liu, Yuqing Song, Yi Liu, Chengjian Qiu, Yangyang Tang, Qiaoying Teng, and Zhe Liu. An effective semi-supervised approach for liver ct image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018.

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.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Rushi Jiao, Yichi Zhang, Le Ding, Rong Cai, and Jicong Zhang. Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *arXiv preprint arXiv:2207.14191*, 2022.

Tom Joy, Sebastian M Schmon, Philip HS Torr, N Siddharth, and Tom Rainforth. Rethinking semi-supervised learning in vaes. *arXiv preprint arXiv:2006.10102*, 2020.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pages 429–445. Springer, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Tom Fletcher. Semi-supervised learning with gans: Manifold invariance with improved inference. *Advances in neural information processing systems*, 30, 2017.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

Chongxuan Li, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. *Advances in neural information processing systems*, 30, 2017.

Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8300–8311, 2021.

Manhua Liu, Fan Li, Hao Yan, Kundong Wang, Yixin Ma, Li Shen, Mingqing Xu, Alzheimer's Disease Neuroimaging Initiative, et al. A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease. *Neuroimage*, 208: 116459, 2020.

Sijie Liu, Jingmin Xin, Jiayi Wu, and Peiwen Shi. Semi-supervised adversarial learning for diabetic retinopathy screening. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 60–68. Springer, 2019.

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1485–1488, 2010.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 39(8): 2688–2700, 2020.

Ozan Oktay, Jay Nanavati, Anton Schwaighofer, David Carter, Melissa Bristow, Ryutaro Tanno, Rajesh Jena, Gill Barnett, David Noble, Yvonne Rimmer, et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. *JAMA network open*, 3(11): e2027426–e2027426, 2020.

Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

Himashi Peiris, Zhaolin Chen, Gary Egan, and Mehrtash Harandi. Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–438. Springer, 2021.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a SyleGAN encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

David Robben, Anna MM Boers, Henk A Marquering, Lucianne LCM Langezaal, Yvo BWEM Roos, Robert J van Oostenbrugge, Wim H van Zwam, Diederik WJ Dippel, Charles BLM Majoie, Aad van der Lugt, et al. Prediction of final infarct volume from native CT perfusion and treatment parameters using deep learning. *Medical image analysis*, 59:101589, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Ali Reza Sajun and Imran Zualkernan. Survey on implementations of generative adversarial networks for semi-supervised learning. *Applied Sciences*, 12(3):1718, 2022.

Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in neural information processing systems*, 21, 2008.

You-Bao Tang, Yu-Xing Tang, Jing Xiao, and Ronald M Summers. Xlsor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *International Conference on Medical Imaging with Deep Learning*, pages 457–467. PMLR, 2019.

Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using

supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *International Conference on Learning Representations*, 2018.

Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, and Hau-San Wong. Enhancing triplegan for semi-supervised conditional instance synthesis and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10091–10100, 2019.

Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Zhehao Zhang, Xiaobo Zhang, Dengmin Wen, Lilan Peng, and Yuxin Zhou. A novel semi-supervised neural network for recognizing parkinson's disease. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–130. Springer, 2022.

## Appendix A. Training details

We resize the images to $256 \times 256$ resolution and normalise the pixel values to the range $[-1, 1]$ for both images and segmentation maps to be compatible with StyleGAN's architecture. In line with Oh et al. (2020) prior to rescaling, we equalise the intensity histogram and apply a gamma correction of factor 0.5.

For SemanticGAN, we use the same hyperparameters as Li et al. (2021). We train the first step for 100k steps with a batch size of 8 on a single GPU. We choose the model at the step with the lowest FID score, and train the second step for 200k batches with a batch size of 4 (800k sample iterations).

For SemanticAN, we use the same hyperparameters as Li et al. (2021) but a smaller learning rate of $2 \times 10^{-4}$. We train for 50k batches with a batch size of 32 (1,600k sample iterations) and choose the model with the highest validation DICE score.

For SupOnly, we use an Adam optimiser (Kingma and Ba, 2015) with learning rate of $10^{-5}$ and weight decay of $10^{-4}$. We train

Table A.I: Datasets used in our experiments

| Name | Segmentation labels | # scans | # males | # females |
|---|---|---|---|---|
| ChestX-ray8 | No | 108,000 | 60,999 (56%) | 47,001 (44%) |
| JSRT (train) | Yes | 350 | 171 (49%) | 179 (51%) |
| JSRT (test) | Yes | 81 | 35 (43%) | 46 (57%) |
| NIH* | Yes | 95 | 53 (56%) | 42 (44%) |
| Montgomery | Yes | 138 | 63 (46%) | 74 (54%) |
| CheXpert | No | 76,205 | 44,773 (59%) | 31,432 (41%) |

*The NIH dataset is an annotated subset of ChestX-ray8.

for 10k batches with a batch size of 32 (320k sample iterations) and choose the model with the highest validation DICE score.

Throughout the paper, any mention of statistical significance implies that we carried out a dependent t-test for paired variables and an independent Welch's t-test for independent variables, and the p-value was below 0.05.

**Downstream task performance for disease classification** For the classification model, we use a DenseNet (Huang et al., 2017) pre-trained on ImageNet from torchvision (Marcel and Rodriguez, 2010). We replace the last layer with a dense layer and randomised weights. We use the CheXpert dataset, set the intensities of pixels outside the lung area to zero, and separate training and test sets. We experiment with rescaling the scans' intensities between 0 and 1, and find that the model performs better when the intensities are unnormalised – between 0 and 255. We optimise the model with an Adam optimiser with learning rate of $10^{-3}$.

Table A.II: Diagnosis classification AUROC for images masked using different segmentation models on CheXpert (out-of-domain).

| | Opacity | Effusion | Oedema |
|---|---|---|---|
| SupOnly UN | 66.3 | 80.0 | 76.7 |
| SupOnly DL | 67.5 | 82.0 | 78.0 |
| SemanticAN UN | 67.1 | 81.5 | 77.3 |
| SemanticAN DL | 66.8 | 78.8 | 77.3 |
| SemanticGAN | **68.1** | **82.2** | **78.4** |
| Unmasked | **70.3** | **85.3** | **79.0** |

## Appendix B. Comparing U-Net and DeepLabV3 backbones

This appendix chapter aims to compare the performance of DeepLabV3 and U-Net for fully supervised and semi-supervised segmentation. In Experiment 1 – shown in Figure A.III – when only using labelled images, DeepLabV3 outperforms both U-Net and all unsupervised methods in most testing circumstances. However, it is interesting to see that the adversarially trained U-Net performs comparably to the adversarially trained DeepLabV3 for the labelled in-domain data, and better for the unlabelled in-domain data and out-of-domain data. Similarly, in Experiment 2 – shown in Figure A.II – SemanticAN trained with the U-Net backbone outperforms the DeepLabV3-based SemanticAN. These results highlight that DeepLabV3 overfits to the labelled training data.

## Appendix C. Downstream task subgroup performance

Figure A.I shows the stratification per sex of the AUROC for the diagnoses of lung opacity, pleural effusion and oedema using masked images derived in Section 5.2. The difference in the performance of models between males and females was not statistically significant, except for the unmasked image model, which performed better for females than males for oedema and pleural effusion. Moreover, the difference in performance between models was in line with that observed in Section 5.2, where SemanticGAN outperformed SemanticAN and either performed similarly to SupOnly (lung opacity for females, pleural effusion, edema) or better (lung opacity for male), and the unmasked image model outperformed all other methods.

## Appendix D. Consistency over continuous image changes

As a final experiment, we aim to gain a deeper understanding of SemanticGAN from a qualitative perspective by applying it to PCA-generated neighbouring samples, observing both the image reconstruction and the segmentation. We interpolate scans through PCA instead of well-known generative adversarial interpolation methods to avoid the circularity of using a tool to evaluate itself. We proceed by training a PCA model using 3,000 random images from CheXpert balanced across biological sex and ethnicity. We extract the mean PCA image and four neighbouring images at 1 and 2 standard deviations in each direction for each of the first three principal modes. We then use SemanticGAN to predict the lung segmentation and qualitatively examine the gradual change in both the image reconstructions and label maps.

Figure A.II shows SemanticGAN's image reconstruction and segmentation for different PCA components. The first principal component (columns 2-5) reflects the relative closeness of the patient to the X-ray source. In the left-most image, the patient occupies a more significant percentage of space: the space between the image's border and the patient is small, and the lung fields appear larger. Conversely, in the right-most PCA component, the space between the patient and the image borders is more significant, and the patient's lung areas are smaller. SemanticGAN's segmentations reflect the patient closeness, where the left-most segmentation is 9% larger than the mean, and the right-most segmentation is 10% smaller than the mean. For the second component (columns 6-9), the main change is in the patient positioning, from left to right of the field of view, as can be seen from the image margins. Accordingly, the left-most image segmentation is further left than the mean one, and the right-most image segmentation is further right than the mean one. Finally, for the third component (columns 10-13), the size of the lung area increases from left to right. The segmentation grows accordingly: the left-most segmentation is 49% smaller than the mean one, and the right-most segmentation is 39% larger than the mean one. Interestingly, the model's image reconstruction appeared more realistic than the PCA image, providing a good sanity check of the mechanisms underpinning the model's segmentation production. We found that images which were close in PCA space had similar segmentations. Although the shape of the lung area could not be distinguished, SemanticGAN generated segmentations with consistent shapes but varying sizes. An intuitive explanation for this behaviour is that SemanticGAN learns a continuous template

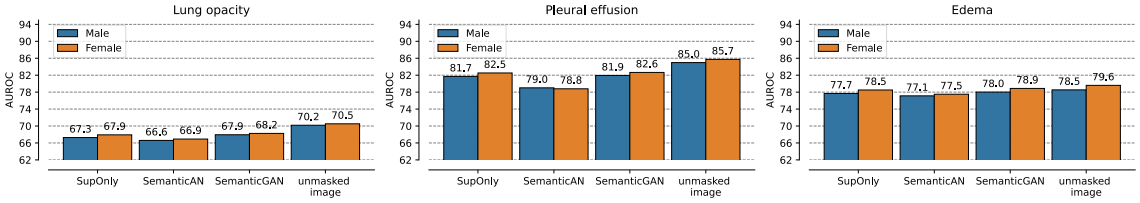



Figure A.I: Disease classification AUROC score for masked images stratified by biological sex.

Table A.III: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. 'UN' stands for U-Net and 'DL' stands for DeepLabV3.

| | Dice | Precision | Recall | Avg. surface distance | Hausdorff distance |
|---|---|---|---|---|---|
| | JSRT (labelled in-domain) | | | | |
| SupOnly UN | 97.3 ± 1.1 | 98.4 ± 1.6 | 96.2 ± 1.7 | 1.4 ± 1.8 | 5.5 ± 9.3 |
| SupOnly DL | **99.3 ± 0.4** | **99.1 ± 0.4** | **99.5 ± 0.5** | **0.3 ± 0.2** | **1.1 ± 0.4** |
| SemanticAN UN | 93.2 ± 2.1 | 89.3 ± 3.5 | 97.6 ± 1.4 | 11.3 ± 4.5 | 54.4 ± 19.0 |
| SemanticAN DL | 91.2 ± 5.2 | 93.3 ± 3.7 | 89.8 ± 7.6 | 3.7 ± 2.0 | 14.7 ± 7.1 |
| SemanticGAN | 98.2 ± 1.3 | 98.3 ± 1.5 | 98.1 ± 1.2 | 0.7 ± 0.6 | 2.5 ± 2.4 |
| | NIH (unlabelled in-domain) | | | | |
| SupOnly UN | 80.2 ± 15.9 | 91.0 ± 15.8 | 72.9 ± 17.2 | 8.6 ± 7.3 | 36.0 ± 24.4 |
| SupOnly DL | **90.3 ± 8.1** | **95.3 ± 4.6** | 86.5 ± 11.4 | **4.8 ± 4.8** | 25.5 ± 22.7 |
| SemanticAN UN | 81.6 ± 9.6 | 76.3 ± 12.6 | 88.8 ± 7.7 | 22.1 ± 7.7 | 71.1 ± 17.6 |
| SemanticAN DL | 72.4 ± 19.6 | 75.4 ± 18.8 | 71.2 ± 21.2 | 10.9 ± 7.13 | 35.5 ± 17.1 |
| SemanticGAN | 90.1 ± 3.4 | 89.9 ± 6.5 | **90.9 ± 5.0** | 4.9 ± 3.3 | **18.2 ± 16.0** |
| | Montgomery (out-of-domain) | | | | |
| SupOnly UN | 92.7 ± 7.9 | 96.6 ± 9.9 | 89.4 ± 6.9 | 3.5 ± 4.7 | 13.7 ± 17.8 |
| SupOnly DL | **96.5 ± 2.3** | **98.8 ± 0.7** | 94.4 ± 4.0 | **1.3 ± 0.7** | **5.5 ± 7.2** |
| SemanticAN UN | 90.3 ± 6.4 | 87.0 ± 8.9 | **94.5 ± 4.0** | 17.6 ± 8.5 | 72.3 ± 27.2 |
| SemanticAN DL | 57.9 ± 33.4 | 62.4 ± 31.0 | 55.5 ± 35.1 | 19.3 ± 15.1 | 54.8 ± 38.9 |
| SemanticGAN | 91.3 ± 5.6 | 92.6 ± 7.9 | 90.4 ± 4.2 | 8.8 ± 10.5 | 34.2 ± 39.1 |

of what the lung segmentation resembles and subsequently adapts it to the given image.

Table A.IV: Detail of training population for Experiment 5.3, where "all" refers to both male and female samples being included.

| Experiment name | $G$ training | | $E$ training | |
|---|---|---|---|---|
| | $\mathcal{D}_u$ | $\mathcal{D}_l$ | $\mathcal{D}_u$ | $\mathcal{D}_l$ |
| Control | all | all | all | all |
| Full bias | males | males | males | males |
| Biased generator $G$ | males | males | all | all |
| Biased encoder $E$ | all | all | males | males |
| Biased labelled dataset $\mathcal{D}_l$ | all | males | all | males |
| Biased unlabelled dataset $\mathcal{D}_u$ | males | all | males | all |

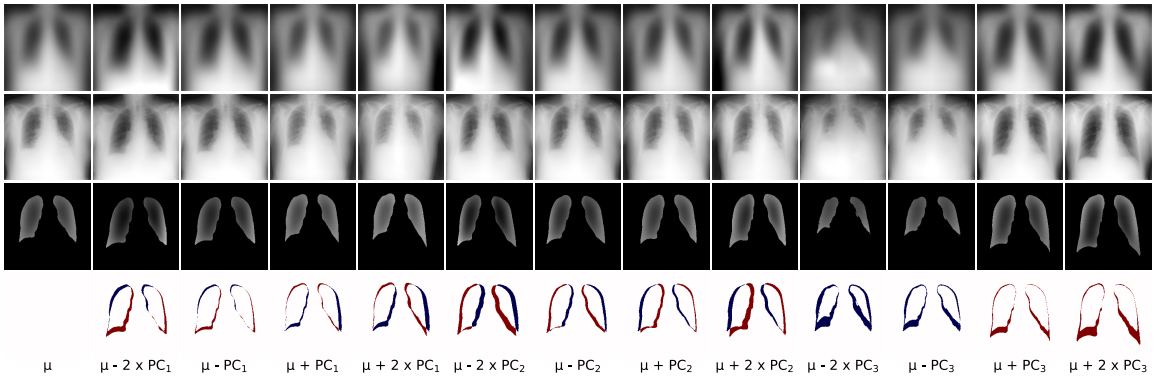


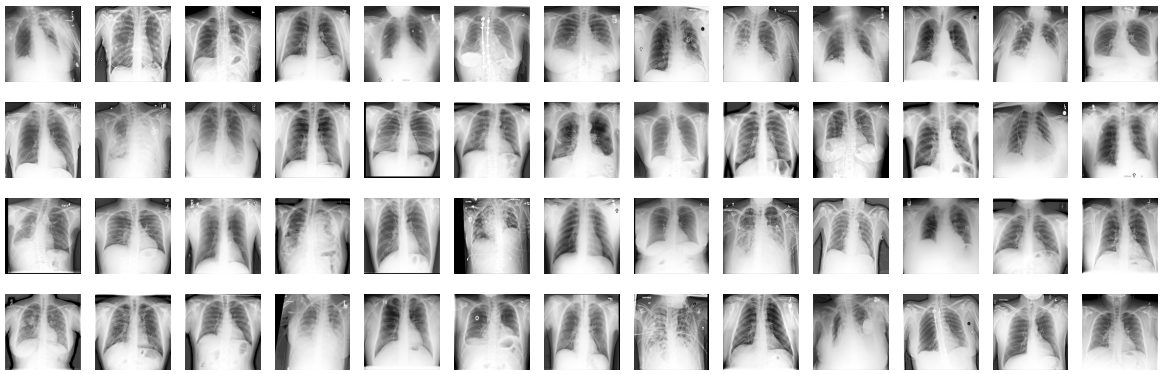

Figure A.II: CheXpert dataset PCA, with generative model reconstruction and segmentation. The first row shows the principal component images, the second row shows SemanticGAN's reconstruction, the third row show its segmentation, and the fourth row shows the difference in segmentation with the mean PCA image.

Table A.V: Stratified models performance for females and males, quantified by Dice similarity coefficient. 'ID' stands for in-domain, 'OOD' stands for out-of-domain.

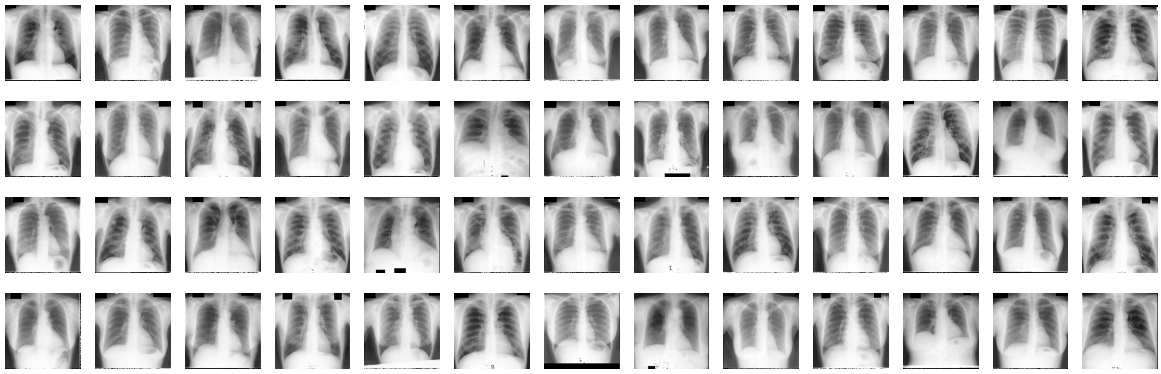
| | Female | Male |
|---|---|---|
| | JSRT (lab. ID) | |
| SupOnly UN | 97.0 ± 1.4 | 97.5 ± 0.5 |
| SupOnly DL | **99.3 ± 0.4** | **99.3 ± 0.4** |
| SemanticAN UN | 92.5 ± 2.6 | 93.9 ± 1.0 |
| SemanticAN DL | 89.9 ± 7.1 | 92.3 ± 1.9 |
| SemanticGAN | 98.2 ± 1.1 | 98.2 ± 1.5 |
| | NIH (unlab. ID) | |
| SupOnly UN | 77.3 ± 18.1 | 82.4 ± 13.8 |
| SupOnly DL | 88.9 ± 9.6 | **91.4 ± 6.6** |
| SemanticAN UN | 79.0 ± 9.8 | 83.6 ± 9.0 |
| SemanticAN DL | 67.5 ± 22.6 | 76.3 ± 16.0 |
| SemanticGAN | 90.5 ± 2.8 | 89.8 ± 3.9 |
| | Montgomery (OOD) | |
| SupOnly UN | 91.8 ± 8.4 | 94.2 ± 5.3 |
| SupOnly DL | 96.5 ± 2.3 | **96.5 ± 2.2** |
| SemanticAN UN | 89.5 ± 5.8 | 91.7 ± 6.2 |
| SemanticAN DL | 44.1 ± 31.2 | 74.8 ± 27.7 |
| SemanticGAN | 90.6 ± 4.5 | 92.3 ± 6.4 |

Table A.VI: Biological sex bias impact on model performance

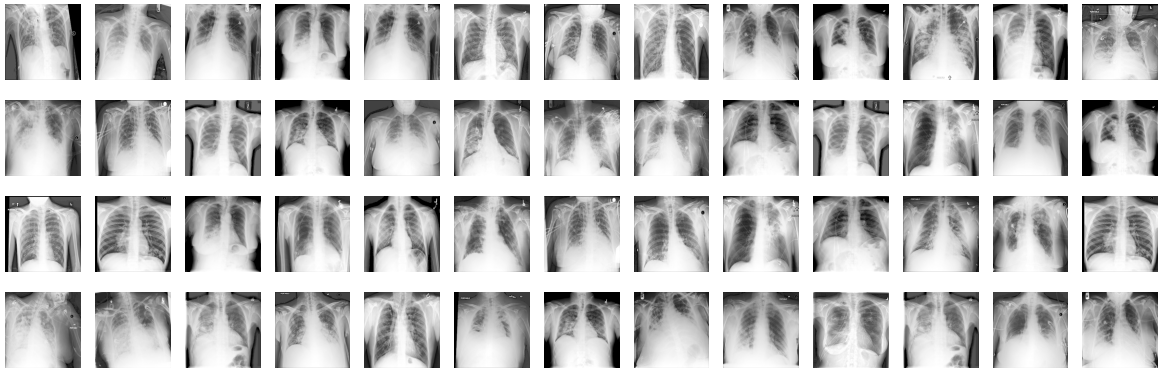
| | Female | Male |
|---|---|---|
| | JSRT (lab. in-domain) | |
| Control | 98.2 ± 1.1 | 98.2 ± 1.5 |
| Full bias | 93.8 ± 2.7 | 98.3 ± 1.6 |
| Biased $G$ | 98.6 ± 0.9 | 98.4 ± 1.4 |
| Biased $E$ | 94.3 ± 1.7 | 98.5 ± 1.6 |
| Biased $\mathcal{D}_l$ | 94.8 ± 1.3 | 98.5 ± 1.6 |
| Biased $\mathcal{D}_u$ | 98.4 ± 1.1 | 98.2 ± 1.7 |
| | NIH (unlab. in-domain) | |
| Control | 90.6 ± 2.8 | 89.8 ± 3.9 |
| Full bias | 89.7 ± 5.1 | 91.0 ± 3.5 |
| Biased $G$ | 92.3 ± 2.8 | 92.3 ± 3.0 |
| Biased $E$ | 88.3 ± 4.2 | 90.0 ± 3.6 |
| Biased $\mathcal{D}_l$ | 90.6 ± 3.5 | 91.2 ± 2.9 |
| Biased $\mathcal{D}_u$ | 88.5 ± 4.6 | 89.9 ± 2.8 |
| | Montgomery (out-of-domain) | |
| Control | 90.6 ± 4.5 | 92.3 ± 6.5 |
| Full bias | 88.4 ± 5.3 | 92.5 ± 5.2 |
| Biased $G$ | 92.4 ± 3.3 | 94.0 ± 3.5 |
| Biased $E$ | 90.3 ± 3.9 | 93.0 ± 3.8 |
| Biased $\mathcal{D}_l$ | 91.2 ± 3.2 | 93.2 ± 3.7 |
| Biased $\mathcal{D}_u$ | 88.8 ± 4.0 | 92.0 ± 5.2 |



(*a*) ChestX-ray8

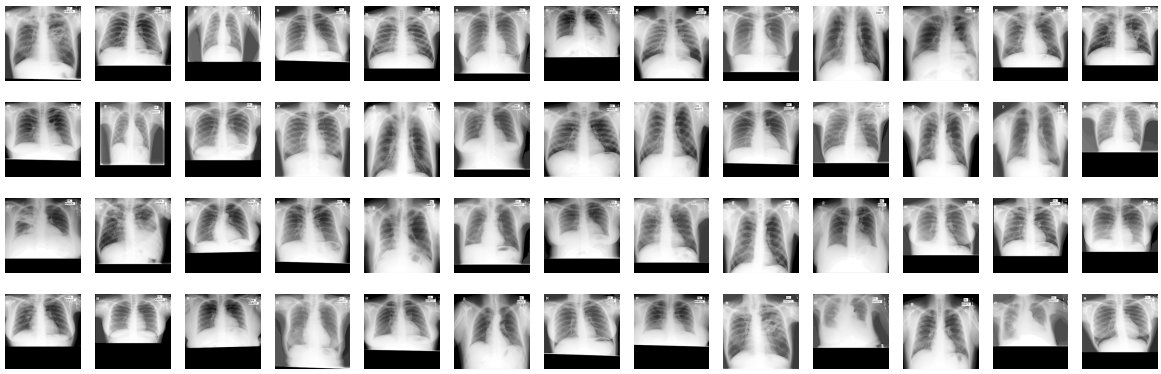


(*b*) JSRT

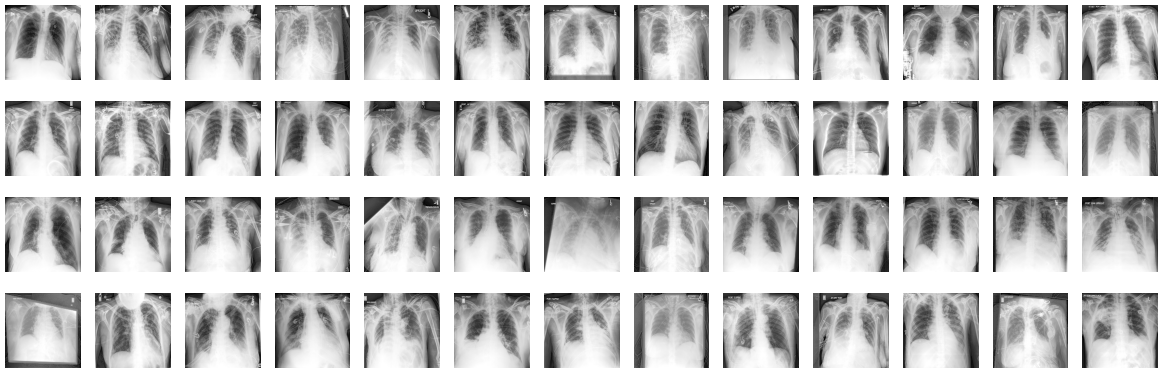


(*c*) NIH

Figure A.III: Datasets thumbnails



(*a*) Montgomery



(*b*) CheXpert

Figure A.IV: Datasets thumbnails - continued