Informative Synthetic Data Generation for Thorax Disease Classification

Yancheng Wang[1] **Rajeev Goel**[1] **Marko Jojic**[1] **Alvin C. Silva**[2] **Teresa Wu**[2] **Yingzhen Yang**[1]

[1]School of Computing and Augmented Intelligence, , Arizona State University
[2]Mayo Clinic Arizona

Abstract

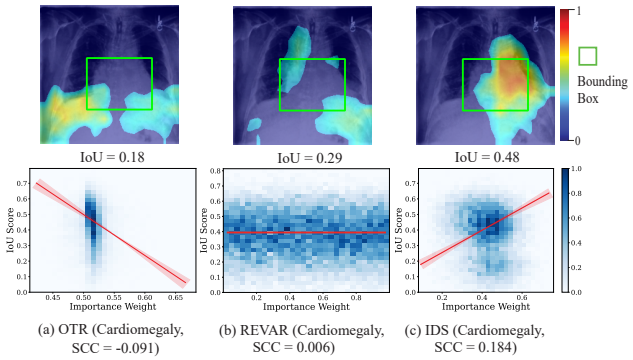
Deep Neural Networks (DNNs), including architectures such as Vision Transformers (ViTs), have achieved remarkable success in medical imaging tasks. However, their performance typically hinges on the availability of large-scale, high-quality labeled datasets—resources that are often scarce or infeasible to obtain in medical domains. Generative Data Augmentation (GDA) offers a promising remedy by supplementing training sets with synthetic data generated via generative models like Diffusion Models (DMs). Yet, this approach introduces a critical challenge: synthetic data often contains significant noise, which can degrade the performance of classifiers trained on such augmented datasets. Prior solutions, including data selection and re-weighting techniques, often rely on access to clean metadata or pretrained external classifiers. In this work, we propose *Informative Data Selection* (IDS), a principled sample re-weighting framework grounded in the Information Bottleneck (IB) principle. IDS assigns higher weights to more informative synthetic samples, thereby improving classifier performance in GDA-enhanced training for thorax disease classification. Extensive experiments demonstrate that IDS significantly outperforms existing data selection and re-weighting baselines. Our code is publicly available at `https://github.com/Statistical-Deep-Learning/IDS`.

1 INTRODUCTION

Recent advances have significantly propelled the use of deep neural networks (DNNs) in medical imaging tasks, particularly for disease classification from chest X-rays [Guendel et al., 2018, Xiao et al., 2023]. Early approaches primarily employed convolutional neural networks (CNNs), such as U-Net [Ronneberger et al., 2015], to facilitate effective representation learning from radiographic data. More recently, Vision Transformers (ViTs) [Dosovitskiy et al., 2020] have been adopted for similar purposes [Xiao et al., 2023], benefiting from their ability to model long-range feature dependencies. Although both CNN- and ViT-based methods have demonstrated promising performance, their success is critically contingent on the availability of high-quality annotated datasets [Feng et al., 2020]. In medical domains, however, acquiring such annotations is often difficult [El Jiani et al., 2022, Xiao et al., 2023] or even infeasible [Esteva et al., 2021, Price and Cohen, 2019, Ali et al., 2023, Ramudu et al., 2023], due to constraints in resources or concerns over data privacy. To mitigate this limitation, self-supervised learning (SSL) approaches, including restorative learning [Xiao et al., 2023], have been explored to extract informative representations from unlabeled data. In parallel, building on the momentum of recent generative modeling breakthroughs [Rombach et al., 2022, Akrout et al., 2023], generative data augmentation (GDA) [Sarıyıldız et al., 2023, Lei et al., 2023, Azizi et al., 2023b, Trabucco et al., 2024a] has emerged as a compelling strategy to synthesize labeled training samples via deep generative models, thereby enhancing the diversity and scale of training datasets.

Generative Data Augmentation (GDA) for Disease Classification. Data scarcity and the absence of high-quality labeled training data have long hindered progress in both medical imaging and general computer vision. To address this limitation, recent work on generative data augmentation (GDA) [Sarıyıldız et al., 2023, Lei et al., 2023, Azizi et al., 2023b, Trabucco et al., 2024a] has explored the use of generative models, including Generative Adversarial Networks (GANs) [Zhang et al., 2021, Li et al., 2022] and Diffusion Models (DMs) [He et al., 2023b, Tian et al., 2023, Yuan et al., 2022, Bansal and Grover, 2023, Vendrow et al., 2023], to synthesize realistic training samples. These approaches have yielded promising outcomes in both general computer vision [Sarıyıldız et al., 2023, Azizi et al., 2023b, Trabucco

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:4532–4557.



(a) OTR (Cardiomegaly, SCC = -0.091) (b) REVAR (Cardiomegaly, SCC = 0.006) (c) IDS (Cardiomegaly, SCC = 0.184)

Figure 1: **Figures in the first row** illustrate examples of thresholded Grad-CAM visualization for OTR, REVAR and IDS. For each of the examples, we also present the ground-truth bounding box for the disease. The thresholded heatmap areas are considered as the disease localization areas. IoU score between the disease localization area and the ground-truth bounding box is shown below each example. **Figures in the second row** illustrate the correlation between IoU scores for disease localization and importance weights for OTR [Guo et al., 2022], REVAR [Jain et al., 2024], and IDS in the CheXpert dataset. The disease name and Spearman Correlation Coefficients (SCC) [Spearman, 1961] are attached in the parenthesis. A larger absolute value of a positive SCC between two variables indicates a stronger positive correlation, which refers to a correlation between two variables where as one variable increases, the other variable tends to increase as well. As a result, a cell with more blue indicates more synthetic images falling in that cell. The red lines in the figures are the linear regression results between the IoU scores and the importance weights, which visualizes the correlation. It is observed that the linear regressors in red suggest a stronger positive correlation between the IoU scores and the importance weights by our IDS than that for competing baselines, which is further quantitatively evidenced by the higher SCC for IDS than the competing baselines. The correlation analysis on NIH ChestX-ray14 is illustrated in Figure 5 in Section D.2 of the supplementary.

et al., 2024a] and medical applications such as image classification [Akrout et al., 2023] and anomaly detection [Wolleb et al., 2022]. Motivated by these successes, this work investigates whether augmenting benchmark thorax disease datasets with synthetic images generated by diffusion models can improve the performance of deep neural networks (DNNs) for thorax disease classification.

Challenges in GDA for Disease Classification. Despite the potential of GDA, synthetic data produced by generative models often exhibit substantial noise [He et al., 2023a, Azizi et al., 2023a], which can negatively impact the performance of classifiers trained on such augmented datasets. To mitigate this, prior studies have employed data selection [Chhabra et al., 2024] or sample re-weighting techniques [He et al., 2023a], where noisy or low-quality synthetic samples are either discarded or down-weighted during training. Sample re-weighting methods [Shu et al., 2019, Guo et al., 2022, Jain et al., 2024] typically rely on training a meta-network using clean metadata to assign higher weights to more informative samples. However, these methods assume access to such metadata, which is often unavailable or impractical to obtain in the medical domain without significant expert involvement. Closest to our problem setting is CBF [He et al., 2023a], which uses a CLIP Filter strategy to remove noisy synthetic images based on the zero-shot classification confidence from the vision-language model CLIP [Radford et al., 2021]. However, CLIP's pretraining on generic image-text pairs may limit its effectiveness on specialized domains such as thorax X-ray disease classification, undermining its reliability in this setting.

Our Contributions. This work introduces a principled sample re-weighting framework based on the Information Bottleneck (IB), which circumvents the need for clean metadata or external classifiers and delivers state-of-the-art results in GDA for thorax disease classification. Our contributions are as follows. First, we propose IDS, a novel IB-driven re-weighting method, which assigns importance weights to synthetic samples to improve classifier performance on augmented datasets. Unlike prior approaches [Shu et al., 2019, Guo et al., 2022, Jain et al., 2024, Chhabra et al., 2024, He et al., 2023a], IDS is metadata- and classifier-free. Second, we introduce an optimization framework where the re-weighting network minimizes an IB loss by generating importance weights that guide the computation of class centroids in both input and representation spaces. This formulation allows us to derive a separable variational upper bound, termed the VIB, enabling tractable optimization via minibatch SGD. Cross-entropy loss and VIB are jointly optimized to train both the classifier and re-weighting network. Experiments on CheXpert [Irvin et al., 2019], COVIDx [Pavlova et al., 2022], and NIH ChestX-ray14 [Wang et al., 2017] benchmarks show that IDS outperforms existing re-weighting [Shu et al., 2019, Guo et al., 2022, Jain et al., 2024] and selection [He et al., 2023a, Chhabra et al., 2024] approaches. Finally, we analyze the correlation between the importance weights and Intersection over Union (IoU) scores for disease localization across baselines and IDS. Higher IoU between the predicted disease region and the ground-truth bounding box indicates more informative samples. As shown in Figure 1, IDS demonstrates a stronger correlation between IoU and learned weights than baselines, validating its effectiveness in prioritizing high-value synthetic data. Further ablation results are detailed in Section 4.3.

2 RELATED WORKS

2.1 MEDICAL IMAGE ANALYSIS WITH DEEP LEARNING

Deep learning has achieved significant advances in photographic image analysis [Lin et al., 2017b,a], driving growing interest in its application to medical imaging. Convolutional neural networks (CNNs), particularly architectures such as U-Net [Falk et al., 2018, Zhou et al., 2018], have laid the foundation for state-of-the-art performance across multiple medical imaging tasks, including image classification [Wang et al., 2019, Ma et al., 2020], object detection [Falk et al., 2019, Yang and Yu, 2021], and semantic segmentation [Yang and Yu, 2021, Yao et al., 2021]. More recently, vision transformers have demonstrated superior performance over CNNs on a wide range of tasks [Zhu et al., 2021, Cai et al., 2023], further advancing the state of the art. Given the challenge of limited annotated medical data, self-supervised learning strategies—especially contrastive learning approaches [Caron et al., 2020, Xiao et al., 2023]—have gained prominence for pre-training models in this domain [Xiao et al., 2023, Chen et al., 2021]. However, unlike photographic images, radiographic images often exhibit high inter-image similarity due to standardized acquisition protocols [Xiang et al., 2021, Haghighi et al., 2022], which poses unique challenges for contrastive learning [He et al., 2020, Chen et al., 2020]. To address these challenges, restorative strategies such as masked autoencoders (MAE) [He et al., 2022] have been employed for pre-training, yielding improvements in representation learning for medical imaging [Xiao et al., 2023].

2.2 INFORMATION BOTTLENECK PRINCIPLE

The Information Bottleneck (IB) principle [Tishby et al., 2000] offers a theoretical framework for understanding generalization in deep neural networks (DNNs). It suggests that an optimal representation should compress input data while preserving task-relevant information, thereby maximizing mutual information with target outputs and minimizing mutual information with inputs. Deep Variational Information Bottleneck (Deep VIB) [Alemi et al., 2017] was the first to incorporate the IB principle into deep learning objectives. Empirical [Lai et al., 2021, Zhou et al., 2022] and theoretical [Kawaguchi et al., 2023] studies confirm that networks better aligned with the IB principle tend to exhibit stronger performance and generalization. Within the medical imaging literature, IB has been widely adopted to guide learning of task-discriminative representations [Wang et al., 2023, Schott et al., 2024, Li et al., 2023]. While most of these works leverage IB for enhancing representation learning in DNNs, our work is distinct in that it employs the IB principle to guide the selection of high-quality synthetic samples for data augmentation in medical image classification—a novel application of the IB framework in this context.

2.3 GENERATIVE DATA AUGMENTATION, DATA SELECTION, AND SAMPLE RE-WEIGHTING

Generative data augmentation (GDA)—the process of generating synthetic samples to improve model training—has emerged as a vital yet challenging topic in deep learning. Recent studies [Sarıyıldız et al., 2023, Lei et al., 2023, Azizi et al., 2023b, Trabucco et al., 2024a] have employed deep generative models [He et al., 2023b, Tian et al., 2023, Yuan et al., 2022, Bansal and Grover, 2023, Vendrow et al., 2023] to synthesize realistic and diverse training data. In the medical imaging domain, GDA has similarly been adopted to alleviate annotation scarcity [Jiang et al., 2018, Sharma and Hamarneh, 2019, Cha et al., 2020, Akrout et al., 2023, Shin et al., 2018], with several works demonstrating improvements in downstream model performance. However, a major concern with synthetic data is the potential introduction of noise [Azizi et al., 2023b, Trabucco et al., 2024b, Na et al., 2024], which can compromise model accuracy. To address this, recent methods fall into three major categories: (1) improving generative quality via model refinement [Sarıyıldız et al., 2023, Zhou et al., 2023]; (2) data selection, which identifies a high-quality subset of samples from noisy data [Wu et al., 2021, Nguyen et al., 2020, Song et al., 2023, Lin et al., 2023, He et al., 2023a, Chhabra et al., 2024]; and (3) data re-weighting, where samples are assigned importance weights to modulate their influence during training [Mo et al., 2019, Shu et al., 2019, Guo et al., 2022, Jain et al., 2024]. For instance, Classifier-Based Filtering (CBF) [He et al., 2023a] selects synthetic samples based on CLIP zero-shot classification confidence, assuming that high-confidence samples are more likely to be useful. Meanwhile, re-weighting approaches like Meta-Weight-Net [Shu et al., 2019], OTR [Guo et al., 2022], and REVAR [Jain et al., 2024] employ meta-learning to derive adaptive sample weights from clean meta-datasets. Each of these paradigms addresses different aspects of the quality-control challenge in using synthetic data for effective model training.

3 INFORMATIVE DATA SELECTION

Given the original training set $\mathcal{D}_{\text{real}} = \{x_i, y_i\}_{i=1}^{N}$ for Thorax disease classification, we aim to generate synthetic training set $\mathcal{D}_{\text{syn}} = \{\widehat{x}_j, \widehat{y}_j\}_{j=1}^{M}$ with diffusion models and train a classifier on the augmented training set $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}$. To address the adverse impact of noisy synthetic samples in the augmented training set, we introduce *Informative Data Selection* (IDS), a sample re-weighting framework that assigns importance weights to synthetic training examples using a dedicated re-weighting network. This re-weighting network is optimized by minimizing a variational upper bound of the Information Bottleneck (IB) loss com-

puted over the synthetic training data, encouraging higher weights for more informative samples and, consequently, enhancing the performance of the classifier trained on the augmented dataset. Section 3.1 outlines the procedure for generating synthetic training samples using diffusion models. We present the derivation of the variational upper bound of the IB loss in Section 3.2. Finally, Section 3.3 details the joint training procedure of the re-weighting network and the classification network within the IDS framework.

3.1 GENERATING SYNTHETIC TRAINING SAMPLES WITH DIFFUSION MODELS

To generate labeled synthetic training samples, we employ a conditional Latent Diffusion Model (LDM) [Rombach et al., 2022] trained with Classifier-Free Guidance (CFG) [Ho and Salimans, 2022] on latent representations of training images. These latent features are extracted using a pre-trained variational autoencoder (VAE) encoder v_{e} from Stable Diffusion [Rombach et al., 2022], and the reconstruction is performed via its decoder v_{d}. As detailed in Section B.1 of the supplementary material, we use Diffusion Transformers (DiTs) [Peebles and Xie, 2023] as the backbone architecture for the LDM. Let $\{h_i\}_{i=1}^N$ denote the latent representations of the real training dataset $\mathcal{D}_{\text{real}}$, where $h_i = v_{\text{e}}(x_i)$ for image x_i. The LDM, parameterized by ω, is trained on the labeled latent set $\{h_i, y_i\}_{i=1}^N$ to minimize the loss $\mathcal{L}_{\text{LDM}}$ defined in Equation (14) of Section B.1. The detailed training procedure is provided in Algorithm 1 in the supplementary.

After training the LDM, we generate a set of latent features $\{\widehat{h}_j\}_{j=1}^M$ corresponding to a predefined label set $\{\widehat{y}_j\}_{j=1}^M$ using the reverse sampling formulation in Equation (13) of Section B. The synthetic images $\{\widehat{x}_j\}_{j=1}^M$ are then reconstructed by decoding the generated latent features through the decoder: $\widehat{x}_j = v_{\text{d}}(\widehat{h}_j)$. In our experiments, the synthetic label set is chosen to match the original class label distribution, i.e., $\{\widehat{y}_j\}_{j=1}^M = \{y_j\}_{j=1}^M$. The full generative process is detailed in Algorithm 2 in the supplementary. The resulting synthetic training dataset $\mathcal{D}_{\text{syn}} = \{\widehat{x}_j, \widehat{y}_j\}_{j=1}^M$ is then combined with the original dataset $\mathcal{D}_{\text{real}}$ to form an augmented dataset $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{syn}}$. This augmented dataset is subsequently used to jointly train the classifier and sample re-weighting network within the IDS framework, as described in Section 3.3.

3.2 VARIATIONAL UPPER BOUND FOR THE IB LOSS

In order to assign higher importance weights to more informative synthetic training samples, we propose to train the re-weighting network by minimizing the IB loss on the synthetic training set. To achieve this goal, we first derive a variational upper bound for the IB loss, which can be optimized by standard SGD algorithms. Given the synthetic training set $\mathcal{D}_{\text{syn}} = \{\widehat{x}_j, \widehat{y}_j\}_{j=1}^M$, we first specify how to compute the IB loss, $\text{IB}(\Theta) = I(\widehat{Z}(\Theta), \widehat{X}) - I(\widehat{Z}(\Theta), \widehat{Y})$, where Θ is the weights of a neural network, $\widehat{X}$ is a random variable representing the input feature of the synthetic training sample, which takes values in $\{\widehat{x}_j\}_{j=1}^M$, $\widehat{Z}(\Theta)$ is a random variable representing the learned feature of the synthetic training sample, which takes values in $\{\widehat{z}_j(\Theta)\}_{j=1}^M$ with $\widehat{z}_j(\Theta)$ being the learned feature for the j-th synthetic training sample. $\widehat{Y}$ is a random variable representing the synthetic class label, which takes values in $\{y_j\}_{j=1}^n$. We define $\mathcal{C}(\theta, \Theta) = \left\{ \left\{ c_k^{(\text{input})}(\theta) \right\}_{k=1}^C, \left\{ c_k^{(\text{feat})}(\theta, \Theta) \right\}_{k=1}^C \right\}$ as the class centroids of the input features and the learned features on the synthetic training set, where θ denotes the parameters of the sample re-weighting network. The formulas for the computation of $\mathcal{C}(\theta, \Theta)$ can be found in Equation (2). We abbreviate $\widehat{Z}(\Theta)$ as $\widehat{Z}$, $c_k^{(\text{input})}(\theta)$ as $c_k^{(\text{input})}$, and $c_k^{(\text{feat})}(\theta, \Theta)$ as $c_k^{(\text{feat})}$ for simplicity of the notations. Then we define the probability that $\widehat{z}_j$ belongs to class a as $\Pr\left[\widehat{Z} \in a\right] = \frac{1}{M} \sum_{j=1}^M \phi(\widehat{z}_j, c_a^{(\text{feat})})$ with $\phi(\widehat{z}_j, c_a^{(\text{feat})}) = \frac{\exp\left(-\left\|\widehat{z}_j - c_a^{(\text{feat})}\right\|_2^2\right)}{\sum_{a=1}^C \exp\left(-\left\|\widehat{z}_j - c_a^{(\text{feat})}\right\|_2^2\right)}$. Similarly, we define the probability that $\widehat{x}_j$ belongs to class b as $\Pr\left[\widehat{X} \in b\right] = \frac{1}{n} \sum_{j=1}^M \phi(x_j, c_b^{(\text{input})})$. Moreover, we have the joint probabilities $\Pr\left[\widehat{Z} \in a, \widehat{X} \in b\right] = \frac{1}{M} \sum_{j=1}^M \phi(\widehat{z}_j, c_a^{(\text{feat})}) \phi(\widehat{x}_j, c_b^{(\text{input})})$ and $\Pr\left[\widehat{Z} \in a, \widehat{Y} = y\right] = \frac{1}{M} \sum_{j=1}^M \phi(\widehat{z}_j, c_a^{(\text{feat})}) \mathbb{I}_{\{\widehat{y}_i = y\}}$ where $\mathbb{I}_{\{\}}$ is an indicator function. As a result, we can compute the mutual information $I(\widehat{Z}, \widehat{X}) = \sum_{a=1}^C \sum_{b=1}^C \Pr\left[\widehat{Z} \in a, \widehat{X} \in b\right] \log \frac{\Pr\left[\widehat{Z} \in a, X \in b\right]}{\Pr\left[\widehat{Z} \in a\right] \Pr\left[\widehat{X} \in b\right]}$, $I(\widehat{Z}, \widehat{Y}) = \sum_{a=1}^C \sum_{y=1}^C \Pr\left[\widehat{Z} \in a, \widehat{Y} = y\right] \log \frac{\Pr\left[\widehat{Z} \in a, \widehat{Y} = y\right]}{\Pr\left[\widehat{Z} \in a\right] \Pr\left[\widehat{Y} = y\right]}$, and then compute the IB loss $\text{IB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}})$. Given a variational distribution $Q(\widehat{Z} \in a | Y = y)$ for $y \in \{1, \ldots, C\}$ and $a \in \{1, \ldots, C\}$, the following theorem gives a variational upper bound, $\text{VIB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}})$, for the IB loss $\text{IB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}})$.

Theorem 3.1.

$$\text{IB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}}) \leq \text{VIB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}}), \quad (1)$$

where

$$\mathrm{VIB}(\mathcal{C}(\theta,\Theta),\Theta,\mathcal{D}_{\text{syn}}) \coloneqq \frac{1}{M}\sum_{j=1}^{M}\mathrm{VIB}(\mathcal{C}(\theta,\Theta),\Theta,\widehat{x}_j),$$

$$\begin{aligned}&\mathrm{VIB}(\mathcal{C}(\theta,\Theta),\Theta,\widehat{x}_j)\\ &\coloneqq \sum_{a=1}^{C}\sum_{b=1}^{C}\phi(\widehat{z}_j,c_a^{(\text{feat})})\phi(\widehat{x}_j,c_b^{(\text{input})})\log\phi(\widehat{x}_j,c_b^{(\text{input})})\\ &-\sum_{a=1}^{C}\sum_{y=1}^{C}\phi(\widehat{z}_j,c_a^{(\text{feat})})\mathbb{I}_{\{\widehat{y}_j=y\}}\log Q(\widehat{Z}\in a|Y=y).\end{aligned}$$

$\mathrm{VIB}(\mathcal{C}(\theta,\Theta),\Theta,\widehat{x}_j)$ can be interpreted as the information bottleneck upper bound for the j-th synthetic image. The proof of this theorem follows by applying Lemma A.1 and Lemma A.2 in Section A of the supplementary. We remark that $\mathrm{VIB}(\Theta)$ is ready to be optimized by standard SGD algorithms because it is separable and expressed as the summation of losses on individual training points. In order to compute $\mathrm{VIB}(\Theta)$ before a new epoch starts, we need to update the variational distribution $Q^{(t)}$ at the end of the previous epoch.

3.3 FORMULATION OF INFORMATIVE DATA SELECTION (IDS)

Given the original training set $\mathcal{D}_{\text{real}}=\{x_i,y_i\}_{i=1}^{N}$ and the synthetic training set $\mathcal{D}_{\text{syn}}=\{\widehat{x}_j,\widehat{y}_j\}_{j=1}^{M}$ generated by the diffusion model, our goal is to train an image classifier $f_\Theta(\cdot)$ on the augmented dataset $\mathcal{D}_{\text{aug}}=\mathcal{D}_{\text{real}}\cup\mathcal{D}_{\text{syn}}$, where $f_\Theta(\cdot)$ denotes a deep neural network (DNN) with parameters Θ. However, naively training the classifier on $\mathcal{D}_{\text{aug}}$ may degrade performance due to the substantial noise potentially present in synthetic samples from $\mathcal{D}_{\text{syn}}$. To mitigate this, we introduce a sample re-weighting network $g_\theta(\cdot)$ that learns importance weights $\{g_\theta(\widehat{x}_j)\in[0,1]\}_{j=1}^{M}$ for the synthetic training instances. Here, $g_\theta(\cdot)$ is also a DNN, with parameters θ. The re-weighting network serves a role analogous to that of the meta-networks employed in prior work [Shu et al., 2019, Jain et al., 2024], which aim to assign training weights based on sample informativeness.

To ensure that $g_\theta(\cdot)$ assigns higher weights to more informative synthetic examples in $\mathcal{D}_{\text{syn}}$, we optimize it via the variational upper bound of the Information Bottleneck (IB) loss, denoted as VIB, computed over $\mathcal{D}_{\text{syn}}$. A critical step in evaluating the VIB involves estimating class centroids in both the input feature space and the latent representation space, using all samples in the augmented training set $\mathcal{D}_{\text{aug}}$. Let $f'_\Theta(\cdot)$ denote the representation backbone of the classifier $f_\Theta(\cdot)$, i.e., the network excluding its final linear layer. These centroids serving as anchors to measure the relevance and compression terms in the IB objective, essential for computing the VIB loss effectively, are computed by

$$\begin{aligned}c_k^{(\text{input})}(\theta)&=\frac{\sum_{i=1}^{N}x_i\mathbb{I}_{\{y_i=k\}}+\sum_{j=1}^{M}g_\theta(\widehat{x}_j)\widehat{x}_j\mathbb{I}_{\{\widehat{y}_j=k\}}}{\sum_{i=1}^{N}\mathbb{I}_{\{y_i=k\}}+\sum_{j=1}^{M}g_\theta(\widehat{x}_j)\mathbb{I}_{\{\widehat{y}_j=k\}}},\\ c_k^{(\text{feat})}(\theta,\Theta)&=\frac{\sum_{i=1}^{N}x_i\mathbb{I}_{\{y_i=k\}}+\sum_{j=1}^{M}g_\theta(\widehat{x}_j)f'_\Theta(\widehat{x}_j)\mathbb{I}_{\{\widehat{y}_j=k\}}}{\sum_{i=1}^{N}\mathbb{I}_{\{y_i=k\}}+\sum_{j=1}^{M}g_\theta(\widehat{x}_j)\mathbb{I}_{\{\widehat{y}_j=k\}}},\end{aligned}\tag{2}$$

where $k\in[C]$ is the class index and C is the number of classes. $\mathbb{I}_{\{\}}$ is an indicator function. Next, the VIB on the synthetic training set $\mathcal{D}_{\text{syn}}$ can be computed using Equation (2). With the sample re-weighting network $g_\theta(\cdot)$, the overall training loss for the classifier $f_\Theta(\cdot)$ on the augmented training set $\mathcal{D}_{\text{aug}}$ is $\mathcal{L}_{\text{train}}(\theta,\Theta,\mathcal{D}_{\text{aug}}) = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CE}\left(f_\Theta(x_i),y_i\right) + \frac{1}{M}\sum_{j=1}^{M}g_\theta(\widehat{x}_j)\mathrm{CE}\left(f_\Theta(\widehat{x}_j),\widehat{y}_j\right)$, where $\mathrm{CE}(,)$ is the cross-entropy function. To train the classifier $f_\Theta(\cdot)$ by minimizing $\mathcal{L}_{\text{train}}(\theta,\Theta,\mathcal{D}_{\text{aug}})$ while training the sample re-weighting network g_θ by minimizing $\mathrm{VIB}(\theta,\Theta,\mathcal{D}_{\text{syn}})$, we formulate a bi-level optimization objective for IDS as

$$\begin{aligned}\Theta^*&=\arg\min_{\Theta}\mathcal{L}_{\text{train}}(\theta^*,\Theta,\mathcal{D}_{\text{aug}}),\\ \text{s.t. }\theta^*&=\arg\min_{\theta}\mathrm{VIB}(\mathcal{C}(\theta,\Theta^*),\Theta^*,\mathcal{D}_{\text{syn}}),\end{aligned}\tag{3}$$

where Θ^* and θ^* are the optimal parameters for the classifier $f_\Theta(\cdot)$ and the sample re-weighting network $g_\theta(\cdot)$. It is worthwhile to emphasize that the re-weighting is performed only on the synthetic data in (2). As mentioned in Section 4.1, the re-weighting can be applied to both real data $\mathcal{D}_{\text{real}}$ and the synthetic data for even better performance shown in Section 4.

Optimization of IDS. To train the classifier $f_\Theta(\cdot)$ and the sample re-weighting network $g_\theta(\cdot)$ under the bi-level objective in Equation (3), we employ an alternating stochastic gradient descent strategy commonly used in bi-level optimization problems [Shu et al., 2019, Algan and Ulusoy, 2021, Jain et al., 2024]. This approach alternates between updating the parameters of the sample re-weighting network and those of the classifier, enabling efficient handling of the dependency between the two learning processes. In this framework, the lower-level optimization aims to learn a sample re-weighting network that assigns importance weights to training samples, which are then used to guide the upper-level optimization of the classifier toward better generalization performance. At the t-th epoch, we first update the re-weighting network parameters by $\theta^{(t)}=\theta^{(t-1)}-\eta_\theta\nabla_\theta\mathrm{VIB}(\mathcal{C}(\theta,\Theta^{(t-1)}),\Theta^{(t-1)},\mathcal{D}_{\text{syn}})$, where η_θ is the learning rate for the sample re-weighting network. Subsequently, the classifier parameters are updated using $\Theta^{(t)}=\Theta^{(t-1)}-\eta_\Theta\nabla_\Theta\mathcal{L}_{\text{train}}(\theta^{(t-1)},\Theta,\mathcal{D}_{\text{aug}})$, where η_Θ is the learning rate for the classifier. Both VIB and $\mathcal{L}_{\text{train}}$ are separable and conducive to mini-batch SGD, allowing the entire training procedure to scale efficiently. The full training algorithm for IDS is summarized in Algorithm 3 in Section C of the supplementary.

We further remark that IDS naturally extends to multi-label classification tasks. Let L denote the number of labels. For each synthetic training sample $\widehat{x}_j \in \mathcal{D}_{\text{syn}}$, the sample re-weighting network outputs a vector of importance weights $g_\theta(\widehat{x}_j) \in [0,1]^L$, where the l-th entry corresponds to the importance of $\widehat{x}_j$ with respect to label l. Both the training loss $\mathcal{L}_{\text{train}}(\theta, \Theta, \mathcal{D}_{\text{aug}})$ and the variational information bottleneck $\text{VIB}(\mathcal{C}(\theta, \Theta), \Theta, \widehat{x}_j)$ are computed separately for each label, and they are denoted as $\mathcal{L}_{\text{train}}(\theta, \Theta, \mathcal{D}_{\text{aug}}, l)$ and $\text{VIB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}}, l)$ for the l-th label. The bi-level optimization in Equation (3) is then modified by replacing the training loss and VIB with their averaged forms, $\frac{1}{L}\sum_{l=1}^{L} \mathcal{L}_{\text{train}}(\theta, \Theta, \mathcal{D}_{\text{aug}}, l)$ and $\frac{1}{L}\sum_{l=1}^{L} \text{VIB}(\mathcal{C}(\theta, \Theta), \Theta, \mathcal{D}_{\text{syn}}, l)$. Such formulation allows IDS to scale to complex multi-label scenarios common in medical imaging while maintaining its theoretical grounding and practical efficiency.

4 EXPERIMENTS

In this section, we present a comprehensive evaluation of our proposed Informative Data Selection (IDS) method across several medical imaging datasets. First, in Section 4.1, we describe the implementation details of our experiments. We compare IDS against other data selection and sample re-weighting techniques on CheXpert, COVIDx, and NIH-ChestXray-14 in Section 4.2. An ablation study analyzing the correlation between disease localization performance and importance weights for IDS and baseline methods is provided in Section 4.3. Details regarding the generation of synthetic images using diffusion models are deferred to Section B.2 of the supplementary. Additional experimental results are available in Section D of the supplementary, with further implementation details and experimental setups described in Section D.1. Additional results from the ablation study are presented in Section D.2 of the supplementary. The statistical significance of IDS's performance improvement over competing baselines is assessed in Section D.3 of the supplementary. Section D.4 of the supplementary also includes an ablation of IDS components and an analysis of training time. In Section D.5, we evaluate the impact of the diffusion model employed for data generation in IDS and analyze the efficiency of the generation process. Section D.6 of the supplementary compares IDS with active learning methods for identifying informative synthetic data. Finally, in Section D.7, we provide additional comparisons with baseline methods for thorax disease classification across the three benchmarks, and in Section D.8, we show Grad-CAM visualization results on the NIH ChestXray-14 dataset.

4.1 IMPLEMENTATION DETAILS

We evaluate the effectiveness of the proposed IDS method for thorax disease classification using two base classification networks, ViT-S and ViT-B [Dosovitskiy et al., 2020], which are pre-trained on $266{,}340$ and $489{,}090$ chest X-rays, respectively, using Masked Autoencoders (MAE) following the setup in [Xiao et al., 2023]. After pre-training, we fine-tune the IDS-augmented networks on three thorax disease classification datasets: CheXpert [Irvin et al., 2019], COVIDx [Pavlova et al., 2022], and NIH ChestX-ray14 [Wang et al., 2017]. Beyond applying IDS for data re-weighting on synthetic data, we further examine its utility for re-weighting both real and synthetic data. Additional implementation details and experimental configurations are deferred to Section D.1 in the supplementary material. For evaluation, we adopt the mean Area Under the Curve (mAUC) as the metric for the multi-label datasets CheXpert and NIH ChestX-ray14, computing mAUC by averaging per-label AUC scores. For the single-label dataset COVIDx, classification accuracy is used as the evaluation metric.

4.2 EXPERIMENTAL RESULTS

CheXpert. Table 1 presents a comparative analysis of IDS with other data selection and data re-weighting methods for GDA on the CheXpert dataset. The baseline ViT-B model achieves an mAUC of 89.3% when fine-tuned directly on CheXpert. When IDS is employed for GDA, the resulting IDS-ViT-B model achieves an improved mAUC of 90.1%, representing a 0.8% gain over the base ViT-B and a 1.1% improvement relative to ViT-B trained with synthetic data. IDS-based models substantially outperform alternative data selection and re-weighting strategies. For instance, IDS-ViT-B surpasses REVAR by 0.8% in mAUC. Furthermore, incorporating IDS to re-weight both real and synthetic data yields additional gains: IDS-ViT-B applied to both data types exceeds the performance of IDS-ViT-B applied only to synthetic data by 0.6% mAUC. These results underscore the strength of IDS in identifying informative samples across both real and synthetic sources. More extensive baseline comparisons are included in Table 8 in Section D.7 of the supplementary. Table 1 compares the performance of competing data selection and data re-weighting methods with our IDS for GDA on CheXpert. The base model ViT-B achieves a mAUC of 89.3% when fine-tuned on the CheXpert dataset. By incorporating IDS for GDA, the IDS-ViT-B model attains a state-of-the-art mAUC of 90.1%, reflecting a 0.8% improvement over the ViT-B and a 1.1% improvement over the ViT-B trained with synthetic data. Notably, IDS models significantly outperform other data selection and data re-weighting methods for GDA. For instance, IDS-ViT-B outperforms REVAR by 0.8% in mAUC. Moreover, applying IDS to re-weight both the real data and the synthetic data further boosts the performance of IDS. For example, IDS-ViT-B re-weighting both the synthetic data and the real data outperforms IDS-ViT-B re-weighting only the synthetic data by 0.6% in mAUC, demonstrating the merits of IDS in selecting informative samples in both

real data and synthetic data. Comparisons with additional baseline methods are provided in Table 8 in Section D.7 of the supplementary.

Table 1: The performance of various state-of-the-art (SOTA) baseline methods on CheXpert. The best results are in bold, and the second-best results are underlined, for each Backbone. Comparisons with more baselines are deferred to Table 8 in Section D.7 of the supplementary. P-values of the t-test between IDS and the best baseline along with their standard deviations for this table, Table 2 and Table 3 are deferred to Table 4 of the supplementary.

Method	Backbone	Atelectasis	Cardiomegaly	Edema	mAUC (%)
MAE [Xiao et al., 2023]	ViT-S/16	83.5	81.8	94.0	89.2
MAE with Synthetic Data		83.0	81.5	94.0	88.6
MW-Net [Shu et al., 2019]		81.7	82.7	94.1	88.9
OTR [Guo et al., 2022]		84.6	81.2	94.2	89.0
IE [Chhabra et al., 2024]		81.7	82.0	94.2	88.9
CBF [He et al., 2023a]		81.4	82.7	94.2	88.8
REVAR [Jain et al., 2024]		83.0	82.7	94.0	89.0
IDS (Ours)		87.5	83.0	94.4	89.6
IDS (Ours, Re-weighting Real Data)		**87.9**	**83.4**	**94.9**	**90.1**
MAE [Xiao et al., 2023]	ViT-B/16	82.7	83.5	93.8	89.3
MAE with Synthetic Data		83.5	82.7	94.0	89.0
MW-Net [Shu et al., 2019]		83.9	82.7	93.8	89.3
OTR [Guo et al., 2022]		85.5	81.6	93.2	89.3
IE [Chhabra et al., 2024]		83.5	82.7	93.8	89.1
CBF [He et al., 2023a]		84.6	81.8	93.8	89.2
REVAR [Jain et al., 2024]		84.0	82.7	93.8	89.3
IDS (Ours)		86.3	84.1	94.7	90.1
IDS (Ours, Re-weighting Real Data)		**86.8**	**84.8**	**95.5**	**90.7**

COVIDx. Table 2 compares IDS with competing methods for GDA on the COVIDx dataset. The baseline ViT-S and ViT-B models, fine-tuned using synthetic data, achieve accuracies of 95.4% and 95.5%, respectively. Applying IDS yields improvements in both models: IDS-ViT-S and IDS-ViT-B achieve accuracy gains of 1.7% and 1.8%, respectively, over their corresponding baselines. IDS-ViT-B sets a new state-of-the-art with an accuracy of 97.3%, which is 1.0% higher than the best-performing prior method, REVAR. Furthermore, re-weighting both real and synthetic data with IDS leads to additional gains: IDS-ViT-B trained with re-weighted real and synthetic data outperforms its counterpart using only re-weighted synthetic data by 0.4% in mAUC. This highlights the value of IDS in extracting signal from both real and synthetic data distributions. Additional baseline comparisons are reported in Table 9 in Section D.7 of the supplementary.

NIH ChestX-ray14. Table 3 presents a comparison between our proposed IDS approach for Group Distributional Alignment (GDA) and several existing data selection and data re-weighting methods on the NIH ChestX-ray14 dataset. This dataset poses a significant challenge for GDA due to its nature as a multi-label classification task with 14 distinct labels. Notably, all competing data selection and re-weighting approaches yield performance that is even worse than the baseline models trained without any synthetic data augmentation. In stark contrast, IDS consistently improves upon the performance of baseline models and achieves significantly better results than alternative data selection and re-weighting strategies. For example, while the base ViT-B model achieves a mean Area Under the Curve (mAUC) of 83.0%, incorporating synthetic data during training without careful selection degrades the performance to 82.1%. Although the application of existing data selection or re-weighting techniques to the synthetic data yields some improvements over this degraded model, their performance still remains inferior to that of the base model trained without synthetic data. On the other hand, IDS-ViT-B not only recovers but exceeds the baseline performance, achieving an mAUC of 83.4%, surpassing the base ViT-B by 0.4%. Furthermore, IDS-ViT-B outperforms REVAR—the strongest among the competing re-weighting baselines—by a margin of 0.9% in mAUC. Applying IDS to re-weight both real and synthetic data results in an additional performance boost; specifically, this dual re-weighting strategy improves the mAUC by 0.5% over using IDS to re-weight only the synthetic data.

Table 2: Performance comparisons between IDS models and SOTA baselines on COVIDx (in accuracy). Comparisons with more baselines are deferred to Table 9 in Section D.7 of the supplementary.

Method	Backbone	Covid-19 Sensitivity	Accuracy
MAE [Xiao et al., 2023]	ViT-S/16	94.5	95.2
MAE with Synthetic Data		98.0	95.4
MW-Net [Shu et al., 2019]		98.1	96.0
OTR [Guo et al., 2022]		98.0	96.2
IE [Chhabra et al., 2024]		98.0	96.0
CBF [He et al., 2023a]		98.4	96.1
REVAR [Jain et al., 2024]		98.2	96.2
IDS (Ours)		98.8	97.1
IDS (Ours, Re-weighting Real Data)		**99.1**	**97.5**
MAE [Xiao et al., 2023]	ViT-B/16	95.5	95.3
MAE with Synthetic Data		98.0	95.5
MW-Net [Shu et al., 2019]		98.5	96.1
OTR [Guo et al., 2022]		98.0	96.1
IE [Chhabra et al., 2024]		98.0	96.0
CBF [He et al., 2023a]		98.1	96.2
REVAR [Jain et al., 2024]		98.2	96.3
IDS (Ours)		99.0	97.3
IDS (Ours, Re-weighting Real Data)		**99.3**	**97.7**

Table 3: Performance comparison between IDS models and SOTA baselines on NIH ChestX-ray14. More baselines are deferred to Table 10 in Section D.7 of the supplementary.

Method	Backbone	mAUC
MAE [Xiao et al., 2023]	ViT-S/16	82.3
MAE with Synthetic Data		81.8
MW-Net [Shu et al., 2019]		82.0
OTR [Guo et al., 2022]		82.0
IE [Chhabra et al., 2024]		82.1
CBF [He et al., 2023a]		82.1
REVAR [Jain et al., 2024]		82.1
IDS (Ours)		82.7
IDS (Ours, Re-weighting Real Data)		**83.2**
MAE [Xiao et al., 2023]	ViT-B/16	83.0
MAE with Synthetic Data		82.1
MW-Net [Shu et al., 2019]		82.3
OTR [Guo et al., 2022]		82.3
IE [Chhabra et al., 2024]		82.5
CBF [He et al., 2023a]		82.5
REVAR [Jain et al., 2024]		82.5
IDS (Ours)		83.4
IDS (Ours, Re-weighting Real Data)		**83.9**

Improvement Significance Analysis. To determine

whether the improvements attained by our IDS method over existing approaches are statistically significant and not attributable to random variation, we conduct controlled experiments using different datasets from Table 1, Table 2, and Table 3. For each method, including IDS and the leading baselines, we perform 10 independent training runs using different random seeds, which govern both the initialization of the neural networks and the splitting of data into training, validation, and test subsets. We then perform a two-sample t-test comparing the distributions of IDS results against those of the best-performing baseline for each dataset. The resulting mean values, standard deviations, and p-values are summarized in Table 4 in Section D.3 of the supplementary material. These statistical tests confirm that the performance gains achieved by IDS are significant, with all p-values satisfying $p \ll 0.05$, indicating that the improvements are highly unlikely to be due to random chance.

4.3 ABLATION STUDY

Study on the Correlation between Disease Localization and Importance Weights. In this section, we predict disease localization areas using Grad-CAM heatmaps [Selvaraju et al., 2017] and evaluate the quality of synthetic images by computing Intersection-over-Union (IoU) scores between the predicted localization areas and the ground-truth disease bounding boxes. Following the methodology of Xiao et al. [2023], the disease localization area for a synthetic image is obtained by thresholding the Grad-CAM heatmap at a fixed value of 0.3 across all experiments. As shown in the illustrative examples in Figure 2, disease localization areas generated by IDS exhibit larger overlaps with ground-truth bounding boxes and yield higher IoU scores compared to competing baselines. To investigate whether more informative synthetic images receive higher importance weights from IDS and competing re-weighting methods, we analyze the correlation between IoU scores and predicted importance weights. Since ground-truth disease bounding boxes are not available for synthetic images, we restrict this study to the Cardiomegaly class, which typically manifests in a consistent anatomical region around the heart in chest X-rays [Amin and Siddiqui, 2019]. We utilize ground-truth bounding boxes for Cardiomegaly from the NIH ChestX-ray14 test set [Wang et al., 2017] as reference bounding boxes in our correlation analysis.

The correlation between individual IoU scores and corresponding importance weights is depicted in the second row of Figure 1, with additional results on the NIH ChestX-ray14 dataset presented in Figure 5 in Section D.2 of the supplementary material. To visualize the trend, we perform linear regression between the IoU scores and the importance weights. The analysis reveals that synthetic images assigned higher importance weights by IDS also tend to have higher IoU scores, suggesting that IDS prioritizes more informative

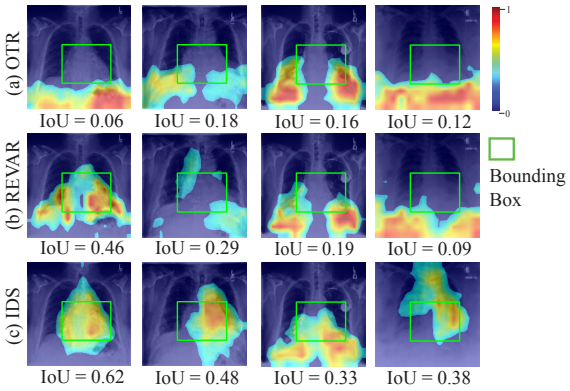


Figure 2: Grad-CAM visualization results on synthetic images for the disease Cardiomegaly from the CheXpert dataset. The Grad-CAM visualizations are shown for (a) OTR, (b) REVAR, and (c) IDS in the first, second, and third rows, respectively. The green boxes represent the ground-truth bounding boxes. These visualizations illustrate that IDS consistently exhibits better disease localization ability compared to OTR [Guo et al., 2022] and REVAR [Jain et al., 2024], as reflected by the higher IoU scores. Grad-CAM visualization results on synthetic images for the disease Cardiomegaly from the NIH ChestX-ray14 dataset are deferred to Figure 4 in Section D.2 of the supplementary.

synthetic samples. In contrast, competing methods such as OTR [Guo et al., 2022] exhibit no positive correlation, while REVAR [Jain et al., 2024] shows only a marginal positive trend. Additionally, we compute the Spearman's rank correlation coefficient (SCC) between individual IoU scores and importance weights. IDS achieves an SCC of 0.184, which substantially exceeds the SCC of 0.006 obtained by REVAR, indicating that IDS more effectively aligns importance weights with the informativeness of synthetic samples.

We further conduct an ablation study to examine the contributions of different components of IDS and report the computational efficiency in Section D.4 of the supplementary material. The results demonstrate the complementary benefits of the variational information bottleneck (VIB) and the re-weighting network for data selection, while maintaining computational feasibility. In addition, the robustness of IDS to different diffusion models is verified in the ablation study presented in Section D.5, indicating that the performance of IDS is not sensitive to the choice of generative backbone. Finally, in Section D.6, we show that IDS significantly outperforms state-of-the-art active learning baselines in identifying informative synthetic data.

5 CONCLUSION

In this paper, we propose Informative Data Selection (IDS), a novel approach for re-weighting synthetic images in Generative Data Augmentation (GDA) by leveraging an

information-theoretic criterion, the Information Bottleneck (IB). IDS optimizes a sample re-weighting network to minimize the IB loss over the synthetic dataset, thereby enforcing the IB principle: learning representations that are more predictive of the output while being minimally dependent on the input. Through comprehensive experiments and ablation analyses, we show that IDS effectively prioritizes more informative synthetic samples in the context of thorax disease classification, and substantially surpasses existing methods in both data selection and re-weighting for GDA.

References

Mohamed Akrout, Bálint Gyepesi, Péter Holló, Adrienn Poór, Blága Kincső, Stephen Solis, Katrina Cirone, Jeremy Kawahara, Dekker Slade, Latif Abid, et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 99–109. Springer, 2023.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Görkem Algan and Ilkay Ulusoy. Meta soft label generation for noisy labels. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7142–7148. IEEE, 2021.

Omar Ali, Wiem Abdelbaki, Anup Shrestha, Ersin Elbasi, Mohammad Abdallah Ali Alryalat, and Yogesh K Dwivedi. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge*, 8(1):100333, 2023.

Imane Allaouzi and Mohamed Ben Ahmed. A novel approach for multi-label chest x-ray classification of common thorax diseases. *IEEE Access*, 7:64279–64288, 2019.

Hina Amin and Waqas J Siddiqui. Cardiomegaly. 2019.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *Trans. Mach. Learn. Res.*, 2023, 2023a.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023b.

Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023.

Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Kenny H Cha, Nicholas Petrick, Aria Pezeshk, Christian G Graff, Diksha Sharma, Andreu Badal, and Berkman Sahiner. Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. *Journal of Medical Imaging*, 7(1):012703–012703, 2020.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Anshuman Chhabra, Peizhao Li, Prasant Mohapatra, and Hongfu Liu. "what data benefits my classifier?" enhancing model performance and interpretability through influence-based data selection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=HE9eUQlAvo`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Laila El Jiani, Sanaa El Filali, et al. Overcome medical image data scarcity by data augmentation techniques: A review. In *2022 International Conference on Microelectronics (ICM)*, pages 21–24. IEEE, 2022.

Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, page 1, 2018.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature methods*, 16(1):67–70, 2019.

Ruibin Feng, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Parts2whole: Self-supervised contrastive learning via reconstruction. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 2020.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision*, pages 510–526. Springer, 2020.

Qingji Guan and Yaping Huang. Multi-label chest x-ray image classification via category-wise residual attention learning. *Pattern Recognition Letters*, 2018.

Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican Congress on Pattern Recognition*, pages 757–765. Springer, 2018.

Dandan Guo, Zhuo Li, He Zhao, Mingyuan Zhou, Hongyuan Zha, et al. Learning to re-weight examples with optimal transport for imbalanced classification. *Advances in Neural Information Processing Systems*, 35: 25517–25530, 2022.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip H. S. Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023b.

Renato Hermoza, Gabriel Maicas, Jacinto C Nascimento, and Gustavo Carneiro. Region proposals for saliency map refinement for weakly-supervised disease localisation and classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–549. Springer, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pages 3–13. Springer, 2021.

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.

Nishant Jain, Karthikeyan Shanmugam, and Pradeep Shenoy. Learning model uncertainty as variance-minimizing instance weights. In *The Twelfth International Conference on Learning Representations*, 2024.

Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Gig S Mageras, Joseph O Deasy, and Harini Veeraraghavan. Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 777–785. Springer, 2018.

Mintong Kang, Yongyi Lu, Alan L Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. *In Submission: Thirty-Sixth Conference on Neural Information Processing Systems*, 2021. URL https://arxiv.org/pdf/2109.12265.pdf.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 16049–16096. PMLR, 2023.

Eunji Kim, Siwon Kim, Minji Seo, and Sungroh Yoon. Xprotonet: Diagnosis in chest radiography with global and local explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15719–15728, June 2021.

Dan Kushnir and Luca Venturi. Diffusion-based sampling for deep active learning. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–9, 2023. doi: 10.1109/SampTA59647.2023.10301392.

Qiuxia Lai, Yu Li, Ailing Zeng, Minhao Liu, Hanqiu Sun, and Qiang Xu. Information bottleneck approach to spatial attention learning. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 779–785. ijcai.org, 2021.

Shiye Lei, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. Image captions are natural prompts for text-to-image models. *arXiv preprint arXiv:2307.08526*, 2023.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.

Guangju Li, Dehu Jin, Qi Yu, and Meng Qi. Ib-transunet: combining information bottleneck and transformer for medical image segmentation. *Journal of King Saud University-Computer and Information Sciences*, 35(3): 249–258, 2023.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8290–8299, 2018.

Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 638–647. IEEE, 2023.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017b.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20706, 2022.

Congbo Ma, Hu Wang, and Steven C. H. Hoi. Multi-label thoracic disease image classification with cross-attention networks, 2020.

Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1378–1382. IEEE, 2019.

Sangwoo Mo, Chiheon Kim, Sungwoong Kim, Minsu Cho, and Jinwoo Shin. Mining gold samples for conditional gans. *Advances in Neural Information Processing Systems*, 32, 2019.

Byeonghu Na, Yeongmin Kim, HeeSun Bae, Jung Hyun Lee, Se Jung Kwon, Wanmo Kang, and Il chul Moon. Label-noise robust diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=HXWTXXtHNl.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. SELF: learning to filter noisy labels with self-ensembling. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Maya Pavlova, Tia Tuinstra, Hossein Aboutalebi, Andy Zhao, Hayden Gunraj, and Alexander Wong. Covidx cxr-3: a large-scale, open-source benchmark dataset of chest x-ray images for computer-aided covid-19 diagnostics. *arXiv preprint arXiv:2206.03671*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Kama Ramudu, V Murali Mohan, D Jyothirmai, DVSSSV Prasad, Ruchi Agrawal, and Sampath Boopathi. Machine learning and artificial intelligence in disease prediction: Applications, challenges, limitations, case studies, and future directions. In *Contemporary Applications of Data Fusion for Advanced Healthcare Informatics*, pages 297–318. IGI Global, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

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.

Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8011–8021, June 2023.

Brayden Schott, Zan Klanecek, Alison Deatsch, Victor Santoro-Fernandes, Thomas Francken, Scott Perlman, and Robert Jeraj. Information bottleneck-based feature weighting for enhanced medical image out-of-distribution detection. In *Submitted to Uncertainty for Safe Utilization of Machine Learning in Medical Imaging - 6th International Workshop*, 2024. URL `https://openreview.net/forum?id=Mshexk31gE`. under review.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 232–243. World Scientific, 2020.

Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39(4):1170–1183, 2019.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer, 2018.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 34(11):8135–8153, 2023. doi: 10.1109/TNNLS.2022.3152527. URL `https://doi.org/10.1109/TNNLS.2022.3152527`.

Charles Spearman. The proof and measurement of association between two things. 1961.

Yuxing Tang, Xiaosong Wang, Adam P Harrison, Le Lu, Jing Xiao, and Ronald M Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018.

Sina Taslimi, Soroush Taslimi, Nima Fathi, Mohammadreza Salehi, and Mohammad Hossein Rohban. Swinchex: Multi-label classification on chest x-ray images with transformers. *arXiv preprint arXiv:2206.04246*, 2022.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024b.

Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation. *arXiv preprint arXiv:2302.07865*, 2023.

Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia. Thoraxnet: an attention regularized deep neural network for classification of thoracic diseases on chest radiography. *IEEE journal of biomedical and health informatics*, 24(2):475–485, 2019.

Junxia Wang, Yuanjie Zheng, Jun Ma, Xinmeng Li, Chongjing Wang, James Gee, Haipeng Wang, and Wenhui Huang. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. *Medical Image Anal.*, 83:102687, 2023. doi: 10.1016/J.MEDIA.2022.102687. URL https://doi.org/10.1016/j.media.2022.102687.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, Nov 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76550-z. URL https://doi.org/10.1038/s41598-020-76550-z.

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.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. NGC: A unified framework for learning with open-world noisy data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 62–71. IEEE, 2021.

Tiange Xiang, Yongyi Liu, Alan L Yuille, Chaoyi Zhang, Weidong Cai, and Zongwei Zhou. In-painting radiography images for unsupervised anomaly detection. *arXiv preprint arXiv:2111.13495*, 2021.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023.

Jianan Yang, Haobo Wang, Sai Wu, Gang Chen, and Junbo Zhao. Towards controlled data augmentations for active learning. In *International Conference on Machine Learning*, pages 39524–39542. PMLR, 2023.

Ruixin Yang and Yingyan Yu. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology*, 11: 638182, 2021.

Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

Yuan Yao, Fengze Liu, Zongwei Zhou, Yan Wang, Wei Shen, Alan Yuille, and Yongyi Lu. Unsupervised domain adaptation through shape modeling for medical image segmentation. In *Medical Imaging with Deep Learning*, 2021.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 93–102, 2019.

Jianhao Yuan, Francesco Pinto, Adam Davies, Aarushi Gupta, and Philip Torr. Not just pretty pictures: Text-to-image generators enable interpretable interventions for robust representations. *arXiv preprint arXiv:2212.11237*, 2022.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez.

Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022.

Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 2018.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021.

Informative Data Selection for Thorax Disease Classification (Supplementary Material)

Yancheng Wang[1] **Rajeev Goel**[1] **Marko Jojic**[1] **Alvin C. Silva**[2] **Teresa Wu**[2] **Yingzhen Yang**[1]

[1]School of Computing and Augmented Intelligence, , Arizona State University
[2]Mayo Clinic Arizona

A PROOF OF THEOREM 3.1

Lemma A.1.

$$I(\widehat{Z}, X) \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\log\phi(x_i, b) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{b=1}^{B}\phi(x_i, b)\log\phi(X_j, b) \tag{4}$$

Proof. By the log sum inequality, we have

$$\begin{aligned}
&I(\widehat{Z}, X)\\
&= \sum_{a=1}^{A}\sum_{b=1}^{B}\Pr\left[\widehat{Z}\in a, X\in b\right]\log\frac{\Pr\left[\widehat{Z}\in a, X\in b\right]}{\Pr\left[\widehat{Z}\in a\right]\Pr\left[X\in b\right]}\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\left(\log\left(\phi(\widehat{z}_j, a)\phi(x_i, b)\right)\right.\\
&\left.-\log\left(\phi(\widehat{z}_j, a)\phi(X_j, b)\right)\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\log\phi(x_i, b)\\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\log\phi(X_j, b)\\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\log\phi(x_i, b)\\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{a=1}^{A}\sum_{b=1}^{B}\phi(\widehat{z}_j, a)\phi(x_i, b)\log\phi(X_j, b).
\end{aligned} \tag{5}$$

□

Lemma A.2.

$$I(\widehat{Z},Y) \geq \frac{1}{n}\sum_{a=1}^{A}\sum_{y=1}^{C}\sum_{i=1}^{n}\phi(\widehat{z}_j,a)\mathbb{I}_{\{y_i=y\}}\log Q(\widehat{Z}\in a|Y=y) \tag{6}$$

Proof. Let $Q(\widehat{Z}|Y)$ be a variational distribution. We have

$$\begin{aligned}
&I(\widehat{Z},Y)\\
&=\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{\Pr\left[\widehat{Z}\in a,Y=y\right]}{\Pr\left[\widehat{Z}\in a\right]\Pr\left[Y=y\right]}\\
&=\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{\Pr\left[\widehat{Z}\in a|Y=y\right]Q(\widehat{Z}\in a|Y=y)}{\Pr\left[\widehat{Z}\in a\right]Q(\widehat{Z}\in a|Y=y)}\\
&\geq\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{\Pr\left[\widehat{Z}\in a|Y=y\right]}{Q(\widehat{Z}\in a|Y=y)}\\
&+\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{Q(\widehat{Z}\in a|Y=y)}{\Pr\left[\widehat{Z}\in a\right]}\\
&=\mathrm{KL}\left(P(\widehat{Z}|Y)\middle\|Q(\widehat{Z}|Y)\right)\\
&+\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{Q(\widehat{Z}\in a|Y=y)}{\Pr\left[\widehat{Z}\in a\right]}\\
&\geq\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log\frac{Q(\widehat{Z}\in a|Y=y)}{\Pr\left[\widehat{Z}\in a\right]}\\
&=\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log Q(\widehat{Z}\in a|Y=y)+H\left(P(\widehat{Z})\right)\\
&\geq\sum_{a=1}^{A}\sum_{y=1}^{C}\Pr\left[\widehat{Z}\in a,Y=y\right]\log Q(\widehat{Z}\in a|Y=y)\\
&\geq\frac{1}{n}\sum_{a=1}^{A}\sum_{y=1}^{C}\sum_{i=1}^{n}\phi(\widehat{z}_j,a)\mathbb{I}_{\{y_i=y\}}\log Q(\widehat{Z}\in a|Y=y).
\end{aligned}\tag{7}$$

□

B INFORMATION ON DIFFUSION MODELS

B.1 FORMULATIONS OF DIFFUSION MODELS

Diffusion models (DMs) are latent variable models that conceptualize data x^0 as a Markov chain progressing from x_T to x^0, with all intermediate variables maintaining consistent dimensions. These models involve two primary Markovian processes: a forward diffusion process defined as $q(x^{(1:T)} \mid x^0) = \prod_{t=1}^{T} q(x^{(t)} \mid x^{(t-1)})$ and a reverse denoising process described by $p_\omega(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\omega(x^{(t-1)} \mid x^{(t)})$. The forward process methodically incorporates Gaussian noise into data $x^{(t)}$:

$$q(x^{(t)} \mid x^{(t-1)}) = \mathcal{N}(x^{(t)}; \sqrt{1-\beta^{(t)}}x^{(t-1)}, \beta^{(t)}\mathbf{I}), \tag{8}$$

where the hyperparameter series $\beta^{(1:T)}$ dictates the noise level added at each step t. The chosen $\beta^{(1:T)}$ ensures that samples x_T approximate standard Gaussian distributions, i.e., $q(x_T) \approx \mathcal{N}(0, \mathbf{I})$. Typically, this forward process q is not adjustable post-definition.

The generation method for DMs involves learning a parameter-driven reverse denoising process to systematically purify the noisy variables $x_{T:1}$ back to the pristine data x^0:

$$p_\omega(x^{(t-1)} \mid x^{(t)}) = \mathcal{N}(x^{(t-1)}; \mu_\omega(x^{(t)}, t), (\rho^{(t)})^2\mathbf{I}), \tag{9}$$

with the initial distribution $p(x_T)$ set as $\mathcal{N}(0, \mathbf{I})$. The model utilizes neural networks like U-Nets or Transformers for calculating means μ_ω, with variances $\rho^{(t)}$ usually predefined.

In terms of optimization, the forward process $q(x^{(1:T)}|x^0)$ is treated as a fixed posterior, against which the reverse process $p_\omega(x_{0:T})$ is trained to enhance the variational lower bound of the data likelihood. Direct likelihood optimization can lead to significant training instability. An alternative simple surrogate objective suggested is:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x^0, \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}), t} \left\| \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\omega(x^{(t)}, t) \right\|_2^2, \tag{10}$$

where the model $\boldsymbol{\varepsilon}_\omega$ predicts the noise vector $\boldsymbol{\varepsilon}$ to clarify diffused samples $x^{(t)}$ at every stage t back to $x^{(t-1)}$. Post-training, samples are generated through iterative ancestral sampling:

$$x^{(t-1)} = \frac{1}{\sqrt{1-\beta^{(t)}}}(x^{(t)} - \frac{\beta^{(t)}}{\sqrt{1-(\alpha^{(t)})^2}}\boldsymbol{\varepsilon}_\omega(x^{(t)}, t)) + \rho^{(t)}\boldsymbol{\varepsilon}, \tag{11}$$

starting from a Gaussian prior $x_T \sim p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$.

Latent Diffusion Models (LDMs) enhance standard Diffusion Models by introducing a latent space that reduces the dimensionality of the data involved in the diffusion process. Initially, data x^0 is encoded to a lower-dimensional latent form h^0. The forward process in LDMs involves:

$$q(h^{(t)} \mid h^{(t-1)}) = \mathcal{N}(h^{(t)}; \sqrt{1-\beta^{(t)}}h^{(t-1)}, \beta^{(t)}I), \tag{12}$$

and the reverse process reconstructs the original clean latent state h^0 from h_T by:

$$p_\omega(h^{(t-1)} \mid h^{(t)}) = \mathcal{N}(h^{(t-1)}; \mu_\omega(h^{(t)}, t), (\rho^{(t)})^2 I), \tag{13}$$

followed by transforming the reconstructed latent data h^0 back to the original data space. The training loss for LDM is

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{h_e(x), \epsilon \sim \mathcal{N}(0, I), t} \left\| \epsilon - \epsilon_\omega(h^{(t)}, t, y) \right\|_2^2, \tag{14}$$

Classifier-Free Guidance (CFG) merges a conditional and an unconditional noise predictor in the sampling process to elevate sample quality and provide class guidance. This technique can be seamlessly integrated into LDMs, formulated as:

$$h^{(t-1)} = \frac{1}{\sqrt{1-\beta^{(t)}}}(h^{(t)} - \frac{\beta^{(t)}}{\sqrt{1-(\alpha^{(t)})^2}}\tilde{\boldsymbol{\varepsilon}}^{(t)}) + \rho^{(t)}\boldsymbol{\varepsilon}, \tag{15}$$

where $\tilde{\boldsymbol{\varepsilon}}^{(t)} = (1+\omega)\boldsymbol{\varepsilon}_\omega(h^{(t)}, y, t) - \gamma\boldsymbol{\varepsilon}_\omega(h^{(t)}, t)$, and γ is the guidance factor, optimizing the sampling process for specific outcomes.

Algorithm 1 describes the training algorithm of the LDM. Algorithm 2 describes the generation process of the synthetic training set.

B.2 DATA GENERATION WITH THE DIFFUSION MODELS

We train the Diffusion Transformer (DiT) on 256×256 images, following the protocol outlined in [Peebles and Xie, 2023]. The training process spans 2,800 epochs with a global batch size of 512, distributed across four NVIDIA A100 GPUs. A constant learning rate of 1×10^{-4} is maintained throughout the training. After training, we generate synthetic images using a classifier-free guidance (CFG) scale of 4.0 with 128 sampling steps. The synthetic dataset is constructed to mirror the

Algorithm 1 Training Algorithm of LDM

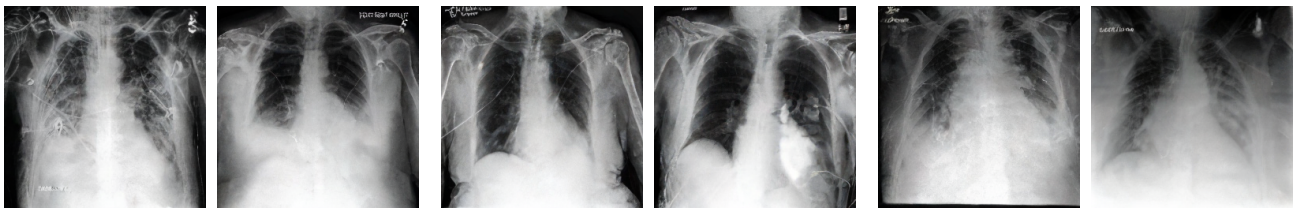
Input: The original training set $\mathcal{D}_{\text{real}} = \{x_i, y_i\}_{i=1}^{N}$, the encoder v_{e} of the fixed pre-trained VAE, and the training epochs of the LDM t_{LDM}.
Output: The parameters of the LDM ω.

1: Initialize the parameter ω of the LDM.
2: Encode input features $\{x_i\}_{i=1}^{N}$ to the latent features $\{h_i\}_{i=1}^{N}$ using the encoder v_{e} such that $h_i = v_{\text{e}}(x_i)$.
3: **for** $t = 1, 2, \ldots, t_{\text{LDM}}$ **do**
4: Update ω by mini-batch SGD on $\{h_i\}_{i=1}^{N}$ using the loss $\mathcal{L}_{\text{LDM}}$ in Equation (14).
5: **end for**
6: **return** The parameters of the LDM ω.

Algorithm 2 Generation of Synthetic Training Set

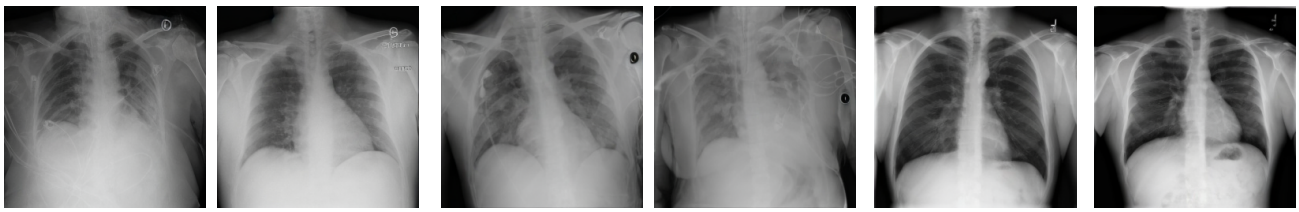
Input: The labels of the synthetic training set $\{\widehat{y}_j\}_{j=1}^{M}$, the parameters of the LDM ω, and the decoder v_{d} of the fixed pre-trained VAE.
Output: The synthetic training set $\mathcal{D}_{\text{syn}} = \{\widehat{x}_i, \widehat{y}_i\}_{j=1}^{M}$.

1: **for** $j = 1, 2, \ldots, M$ **do**
2: Sample a Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$
3: Generate synthetic latent feature $\widehat{h}_j$ from ϵ with the LDM using Equation (13) in Section B of the supplementary.
4: Decode latent feature $\widehat{h}_j$ to the synthetic input feature $\widehat{x}_j$ by $\widehat{x}_j = v_{\text{d}}(\widehat{h}_j)$.
5: **end for**
6: **return** The synthetic training set $\mathcal{D}_{\text{syn}} = \{\widehat{x}_i, \widehat{y}_i\}_{j=1}^{M}$.



(i) Consolidation, Edema and Pleural Effusion (ii) Cardiomegaly and Atelectasis (iii) Cardiomegaly, Edema and Pleural Effusion

(a) CheXpert



(i) COVID-19 (ii) Pneumonia (iii) Normal

(b) COVIDx

Figure 3: Examples of synthetic images generated using a diffusion model trained on the (a) CheXpert and (b) COVIDx datasets, displayed in the first and second rows, respectively. In the first row (CheXpert), the images depict the following medical conditions: (i) Consolidation, Edema, and Pleural Effusion; (ii) Cardiomegaly and Atelectasis; (iii) Cardiomegaly and Pleural Effusion. In the second row (COVIDx), the images correspond to: (i) COVID-19; (ii) Pneumonia; and (iii) Normal (no disease).

label distribution of the real data, ensuring that disease co-occurrence patterns are preserved. Figure 3 presents examples of synthetic images generated by the diffusion model for various thorax diseases. We then integrate these synthetic images into the training sets for COVIDx, CheXpert and NIH-ChestX-ray14. Specifically, we augment the CheXpert, COVIDx and NIH-ChestX-ray14 training sets with $1.0n$ synthetic images, where 'n' denotes the number of images in the official training split of each respective dataset. To ensure fair comparison, all the other baselines are augmented with a similar number of synthetic images.

B.3 COMPUTATION OF $Q^{(t)}(\tilde{\mathbf{x}}|Y)$

The variational distribution $Q^{(t)}(\tilde{\mathbf{x}}|Y)$ can be computed by

$$Q^{(t)}(\widehat{Z} \in a|Y = y) = \Pr\left[\widehat{Z} \in a|Y = y\right] = \frac{\sum_{i=1}^{n} \phi(\widehat{z}_j, a)\mathbb{I}_{\{y_i=y\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{y_i=y\}}}. \tag{16}$$

C ALGORITHM OF IDS

The algorithm for the training process of IDS is described in Algorithm 3.

Algorithm 3 Algorithm of IDS

Input: The augmented training set $\mathcal{D}_{\text{aug}}$, the synthetic training set $\mathcal{D}_{\text{syn}}$, the original training set $\mathcal{D}_{\text{real}}$, epoch number $t_{\max}$.

1: Initialize the classifier network parameters $\Theta^{(0)}$ and the sample re-weighting network parameters $\theta^{(0)}$.
2: **for** $t = 1, 2, \ldots, t_{\max}$ **do**
3: Compute the class centroids of the input features and image representations $\mathcal{C}(\theta, \Theta^{(t-1)})$.
4: Update $\theta^{(t)}$ by applying mini-batch gradient descent on $\mathcal{D}_{\text{syn}}$ using $\theta^{(t)} = \theta^{(t-1)} - \eta_\theta \nabla_\theta \text{VIB}(\mathcal{C}(\theta, \Theta^{(t-1)}), \Theta^{(t-1)}, \mathcal{D}_{\text{syn}})$.
5: Update $\Theta^{(t)}$ byapplying mini-batch gradient descent on $\mathcal{D}_{\text{aug}}$ using $\Theta^{(t)} = \Theta^{(t-1)} - \eta_\Theta \nabla_\Theta \mathcal{L}_{\text{train}}(\theta^{(t-1)}, \Theta, \mathcal{D}_{\text{aug}})$.
6: Compute $Q^{(t)}(\widehat{Z} \in a|\widehat{Y} = y)$ by Eq. (16) in the supplementary.
7: **end for**
8: **return** The trained weights Θ of the classifier network $f_\Theta(\cdot)$ and the trained weights θ of the sample re-weighting network $g_\theta(\cdot)$.

D ADDITIONAL EXPERIMENTS

D.1 ADDITIONAL IMPLEMENTATION DETAILS AND EXPERIMENTAL SETUPS

The fine-tuning process is performed for 75 epochs with the ADAM optimizer and a batch size of 1024. A cosine decay schedule is used. The initial learning rate μ is determined through cross-validation for each model and dataset. The weight decay is set to 0.05, and the momentum parameters β_1 and β_2 are set to 0.9 and 0.999 for all the experiments. We compare our IDS models with several data selection and sample reweighting methods, including Influence Estimation [Chhabra et al., 2024], Classifier-based Filtering (CBF) [He et al., 2023a], MW-Net [Shu et al., 2019], OTR [Guo et al., 2022], and REVAR [Jain et al., 2024]. To ensure a fair comparison, all baseline models undergo an additional 75 epochs of fine-tuning. The mean Area Under the Curve (mAUC) is used as the metric for the multi-label disease classification datasets CheXpert and NIH ChestX-ray14. Accuracy is used as the metric for the single-label disease classification dataset COVIDx.

CheXpert. The CheXpert dataset [Irvin et al., 2019] consists of $224,316$ chest X-ray images from $65,240$ patients, with $191,028$ images used for training. Each X-ray is labeled with radiology reports indicating the presence of 14 thoracic diseases. To measure the effectiveness of our approach, we compute the mean Area Under the Curve (AUC) across five selected disease categories and compare our results against state-of-the-art baseline models.

COVIDx. The COVIDx dataset (Version 9A) [Pavlova et al., 2022] comprises 30,386 chest X-ray images from $17,026$ unique patients. Following the partitioning strategy used in previous studies [Pavlova et al., 2022, Xiao et al., 2023], the dataset is divided into $29,986$ images for training across four classes, and 400 images for testing, categorized into three classes. For objective evaluation and consistency with prior methodologies, we report the Top-1 accuracy on the test set, which contains three classes.

NIH ChestX-ray14. NIH ChestX-ray14 [Wang et al., 2017] is a large-scale dataset comprising $112,120$ chest X-ray images collected from $30,805$ unique patients. Each image may have multiple labels from 14 disease categories, allowing for multi-label classification tasks. Following the official data split provided by Wang et al. [2017], we use $75,312$ images for training and $25,596$ images for testing. The raw images have a resolution of 1024×1024 pixels. In our experiments, we resize the images to 224×224 pixels to match the input requirements of our models. We report the mean Area Under the Curve (AUC) across all 14 disease classes and conduct a comprehensive comparison with 18 widely recognized and influential baseline methods.

D.2 ADDITIONAL STUDY ON THE CORRELATION BETWEEN DISEASE LOCALIZATION AND IMPORTANCE WEIGHTS

Figure 5 illustrates the correlation analysis between IoU scores for disease localization and importance weights on Cardiomegaly for OTR [Guo et al., 2022], REVAR [Jain et al., 2024] and IDS in the NIH-ChestX-ray14 dataset.

As illustrated in Figure 2, the disease localization areas predicted by IDS tend to overlap more with the ground-truth bounding boxes than those predicted by competing baselines, yielding higher IoU scores. To investigate whether IDS assigns higher importance weights to more informative synthetic images, we analyze the correlation between IoU scores and importance weights predicted by IDS and other baseline data re-weighting methods. The second row of Figure 5 illustrates the correlation between individual IoU scores and importance weights. Linear regression is performed to visualize this relationship. The results show that synthetic images assigned higher importance weights by IDS generally have higher IoU scores, indicating that IDS effectively identifies and prioritizes more informative synthetic images. In contrast, there is only a weak positive correlation between importance weights and IoU scores for OTR [Guo et al., 2022] and REVAR [Jain et al., 2024]. To further quantify this correlation, we apply the Spearman Correlation Coefficient (SCC) [Spearman, 1961]. The SCC for IDS is 0.065, significantly higher than the SCC of 0.004 for REVAR, demonstrating that IDS assigns importance weights that are more strongly correlated with IoU scores compared to baseline methods.

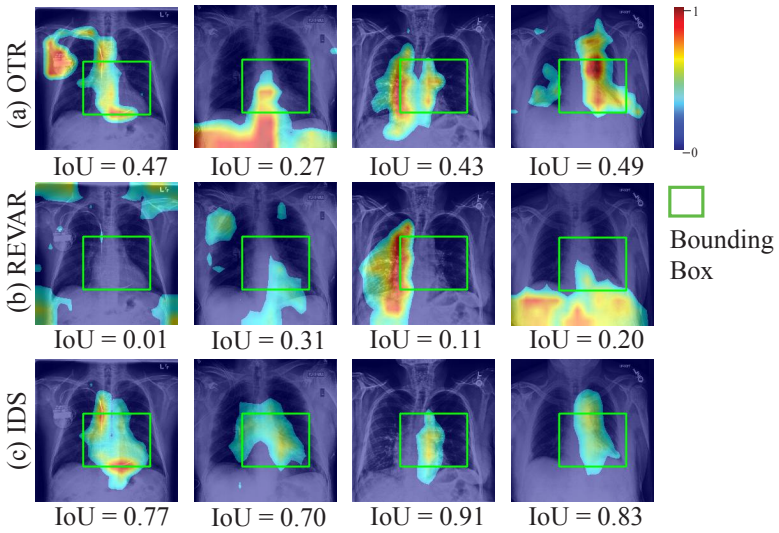
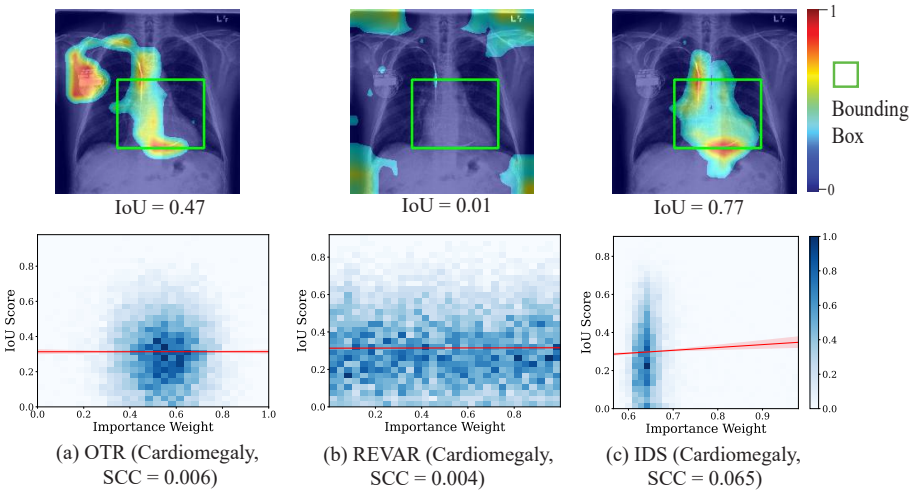


Figure 4: Grad-CAM visualization results on synthetic images for the disease Cardiomegaly from the NiH ChestX-ray14 dataset. The Grad-CAM visualizations are shown for (a) OTR, (b) REVAR, and (c) IDS in the first, second, and third rows, respectively. The green boxes represent the ground-truth bounding boxes. These visualizations illustrate that IDS consistently exhibits better disease localization ability compared to OTR [Guo et al., 2022] and REVAR [Jain et al., 2024], as reflected by the higher IoU scores.

D.3 IMPROVEMENT SIGNIFICANCE ANALYSIS

To verify that the improvement of our proposed IDS on existing methods is statistically significant and out of the range of error, we train both IDS and the best baseline methods on different datasets from Table 1, Table 2, and Table 3 for 10 times with different seeds for random initialization of the networks and train/val/test splits. Next, we perform the t-test between the results of IDS and the results of the best baseline methods on different datasets to assess if the improvement of IDS is statistically significant. The mean and standard deviation of the results and the p-values of the t-test are shown in Table 4. It is observed that the largest p-value is 1.44×10^{-10}, which is less than 0.05. The t-test results suggest that the improvement of IDS over the baseline methods is statistically significant with $p \ll 0.05$, and it is not caused by random error.



(a) OTR (Cardiomegaly, SCC = 0.006) (b) REVAR (Cardiomegaly, SCC = 0.004) (c) IDS (Cardiomegaly, SCC = 0.065)

Figure 5: Figures in the first row are examples of thresholded Grad-CAM visualization for OTR, REVAR, and IDS. For each of the examples, we also present the ground-truth bounding box for the disease Cardiomegaly. The thresholded heatmap areas are considered as the disease localization areas. IoU score between the disease localization area and the ground-truth bounding box is shown below each example. A synthetic image with a higher IoU score is considered a more informative sample for this disease as a larger portion of the predicted disease localization area overlaps with the ground-truth bounding box of the disease. Figures in the second row illustrate the correlation between IoU scores for disease localization and importance weights on Cardiomegaly for OTR [Guo et al., 2022], REVAR [Jain et al., 2024] and IDS in the NIH-ChestX-ray14 dataset. The disease name and Spearman Correlation Coefficients (SCC) [Spearman, 1961] are attached in the parenthesis. A larger absolute value of a positive SCC between two variables indicates a stronger positive correlation, which refers to a correlation between two variables where as one variable increases, the other variable tends to increase as well. The range of IoU and the range of the importance weight, which is $[0, 1] \times [0, 1]$, is divided into 30×30 cells evenly, and the color of each cell is proportional to the number of synthetic images whose IoU sores and importance weights fall in that cell. As a result, a cell with more blue indicates more synthetic images falling in that cell. The red lines in the figures are the linear regression results between the IoU scores and the importance weights, which visualizes the correlation. It can be observed that the linear regressors in red suggest a stronger positive correlation between the IoU scores and the importance weights by our IDS than that for competing baselines, which is further quantitatively evidenced by the higher SCC for IDS than the competing baselines.

Table 4: P-values of t-test between IDS and the best baseline along with their standard deviations on CheXpert, COVIDx, and NIH ChestX-ray14.

Dataset	Backbone	CheXpert (mAUC)	COVIDx (Accuracy)	NIH ChestX-ray14 (mAUC)
Best Baseline	ViT-S/16	89.2 ± 0.067	96.2 ± 0.122	82.3 ± 0.045
IDS		89.6 ± 0.112	97.1 ± 0.125	82.7 ± 0.052
p-value	-	1.44×10^{-10}	3.20×10^{-12}	4.07×10^{-13}
Best Baseline	ViT-B/16	89.3 ± 0.045	96.3 ± 0.158	83.0 ± 0.051
IDS		90.1 ± 0.096	97.3 ± 0.136	83.4 ± 0.065
p-value	-	1.24×10^{-15}	1.40×10^{-11}	1.48×10^{-12}

D.4 ABLATION STUDY AND TRAINING TIME ANALYSIS OF THE IDS

To evaluate the effectiveness and efficiency of different components in the IDS, we compare the disease classification performance and the training time of the baseline model ViT-B, the IDS model IDS-ViT-B, and two ablation models, which are IDS-ViT-B without VIB and IDS-ViT-B without the re-weighting network. The comparison is performed on the COVIDx dataset. The training time is evaluated on four NVIDIA A100 GPUs. The results are shown in Table 5. With only a 30% increase in the training time, IDS-ViT-B improves the classification accuracy on COVIDx by 2.0%, demonstrating the effectiveness of integrating these components into the baseline model. The ablation studies further confirm the individual contributions of the VIB and the re-weighting network, underlining the importance of both components in enhancing model

performance while maintaining a manageable increase in computational demand.

Table 5: Ablation study of IDS with training time analysis. The training time is evaluated on four NVIDIA A100 GPUs.

Methods	COVIDx (Accuracy)	Training Time (minutes/epoch)
ViT-B	95.3	2.6
IDS-ViT-B w/o VIB	96.4	3.2
IDS-ViT-B w/o Re-weighting Network	<u>96.7</u>	2.8
IDS-ViT-B	**97.3**	3.4

D.5 STUDY ON THE DIFFUSION MODELS FOR THE DATA GENERATION IN THE IDS

To evaluate the impact of the diffusion model used for the data generation in the IDS, we compare the performance of IDS-ViT-B using three different diffusion models for data generation, which are DiT-B, DiT-L, and DiT-XL [Peebles and Xie, 2023]. The data generation time and the classification accuracy on the COVIDx dataset are shown in Table 6. It is observed that the performance of the IDS model is not sensitive to the selection of the diffusion models used for data generation. The IDS-ViT-B based on the largest DiT model DiT-XL only outperforms the IDS-ViT-B based on the smallest DiT model DiT-B by 0.2% in classification accuracy on COVIDx, demonstrating the merit of IDS in mitigating the noise in the synthetic data generated by diffusion models. In addition, the results in Table 6 show that the synthetic data generation process with the diffusion models in IDS is efficient, with less than 0.01 seconds/image.

Table 6: Performance comparison between IDS-ViT-B models utilizing different diffusion models for data generation. The data generation time is evaluated on four NVIDIA A100 GPUs.

Methods	COVIDx (Accuracy)	Generation Time (seconds/image)
ViT-B	95.3	-
IDS-ViT-B (DiT-B)	<u>97.1</u>	0.095
IDS-ViT-B (DiT-L)	**97.3**	0.151
IDS-ViT-B (DiT-XL)	**97.3**	0.176

D.6 COMPARISON BETWEEN IDS AND ACTIVE LEARNING METHODS

Active learning (AL) methods aim to minimize the effort required for labeling training data by strategically choosing the most informative instances for annotation [Sinha et al., 2019, Yoo and Kweon, 2019, Gao et al., 2020, Kushnir and Venturi, 2023, Yang et al., 2023, Chhabra et al., 2024]. The selection of the data for annotation by active learning methods is usually achieved by identifying the most informative data points. Such a process works similarly to the data r-weighting process in IDS for identifying the most informative synthetic data. To show the advantage of IDS over active learning methods in selecting the most informative synthetic data, we compare IDS with two state-of-the-art active learning methods, which are CAMPAL [Yang et al., 2023] and SAAL [Chhabra et al., 2024]. Both CAMPAL and SAAL are adopted to select data from the synthetic dataset generated by the diffusion models. The results are shown in Table 7. It is observed that IDS outperforms the competing active learning methods on all the datasets, demonstrating the superiority of IDS in selecting informative training samples compared to active learning methods.

Table 7: Comparison between IDS and active learning methods.

Methods	COVIDx (mAUC)	Covid-19 (Accuracy)	NIH ChestX-ray14 (mAUC)
ViT-B	89.3	95.3	83.0
CAMPAL-ViT-B	<u>89.4</u>	<u>96.2</u>	83.0
SAAL-ViT-B	89.3	95.9	<u>83.1</u>
IDS-ViT-B	**89.6**	**97.3**	**83.4**

D.7 COMPARISON WITH MORE EXISTING WORKS ON THORAX DISEASE CLASSIFICATION

We compare our IDS models with more baselines for thorax disease classification on CheXpert, COVIDx, and NIH-ChestXray-14 in Table 8, Table 9, and Table 10, respectively.

CheXpert. Table 8 presents a performance comparison between additional baseline models and those enhanced by our Informative Data Selection (IDS) technique. For instance, IDS-ViT-B achieves significant improvements, with gains of up to 7.3% in mAUC over the baseline models. In addition to the overall mAUC, Table 8 also provides AUC scores for key thoracic diseases, including Atelectasis, Cardiomegaly, and Edema. These individual disease-specific results further emphasize the effectiveness of IDS, as it consistently boosts performance across various conditions. These findings highlight the superior capabilities of IDS-enhanced models compared to standard baselines on the CheXpert dataset.

COVIDx. Table 9 presents performance comparisons between additional baseline models and our IDS-enhanced models on the COVIDx dataset. For instance, IDS-ViT-B significantly outperforms the baseline models, with accuracy gains of up to 4.7%. Moreover, IDS-ViT-S and IDS-ViT-B achieve a state-of-the-art COVID-19 sensitivity of 99.0%, surpassing previous baselines by up to 11.9%. These results demonstrate the effectiveness of integrating IDS into transformer-based models for medical image analysis on the COVIDx dataset.

NIH-ChestX-ray14. Table 10 compares the performance of various state-of-the-art (SOTA) CNN-based and transformer-based models, including those enhanced by our Informative Data Selection (IDS) technique, on the NIH ChestX-ray14 dataset. The table includes models pre-trained on both ImageNet and X-rays. IDS-ViT-B shows significant improvements, achieving gains of up to 8.9% in mAUC and 8.2% for IDS-ViT-S over baseline models. These gains highlight the effectiveness of IDS in improving performance for thoracic disease classification. Furthermore, Table 10 presents mAUC scores for all methods, demonstrating that IDS-enhanced models consistently outperform other baseline methods, including both CNN and transformer-based Backbones, on the NIH ChestX-ray14 dataset. These findings underscore the superior capabilities of IDS-enhanced models in addressing the challenges of thoracic disease classification.

Table 8: The performance of various state-of-the-art (SOTA) baseline methods on CheXpert. DN represents DenseNet, where the second best performance is underlined.

Method	Backbone	Atelectasis	Cardiomegaly	Edema	mAUC (%)
Allaouzi et al.[Allaouzi and Ahmed, 2019]	DN121	72.0	**88.0**	87.0	82.8
Irvin et al.[Irvin et al., 2019]		81.8	82.8	**93.4**	88.9
Chexclusion [Seyyed-Kalantari et al., 2020]		81.2	83.0	88.3	87.3
Pham et al.[Pham et al., 2021]		**82.5**	85.5	<u>93.0</u>	89.4
BMTL [Hosseinzadeh Taher et al., 2021]		-	-	-	87.1
DiRA [Haghighi et al., 2022]		-	-	-	87.6
Label-assemble [Kang et al., 2021]		<u>82.1</u>	<u>85.9</u>	89.2	89.0
MoCo v2 [Xiao et al., 2023]		78.5	77.9	92.8	88.7
MAE [Xiao et al., 2023]		81.5	77.6	92.3	88.7
MAE [Xiao et al., 2023]	ViT-S/16	83.5	81.8	94.0	<u>89.2</u>
MAE with Synthetic Data		83.0	81.5	94.0	88.6
MW-Net [Shu et al., 2019]		81.7	<u>82.7</u>	94.1	88.9
OTR [Guo et al., 2022]		<u>84.6</u>	81.2	<u>94.2</u>	89.0
IE [Chhabra et al., 2024]		81.7	82.0	<u>94.2</u>	88.9
CBF [He et al., 2023a]		81.4	<u>82.7</u>	<u>94.2</u>	88.8
REVAR [Jain et al., 2024]		83.0	<u>82.7</u>	94.0	89.0
IDS (Ours)		**87.5**	**83.0**	**94.4**	**89.6**
MAE [Xiao et al., 2023]	ViT-B/16	82.7	<u>83.5</u>	93.8	<u>89.3</u>
MAE with Synthetic Data		83.5	82.7	<u>94.0</u>	89.0
MW-Net [Shu et al., 2019]		83.9	82.7	93.8	<u>89.3</u>
OTR [Guo et al., 2022]		85.5	81.6	93.2	<u>89.3</u>
IE [Chhabra et al., 2024]		83.5	82.7	93.8	89.1
CBF [He et al., 2023a]		84.6	81.8	93.8	89.2
REVAR [Jain et al., 2024]		84.0	82.7	93.8	<u>89.3</u>
IDS (Ours)		**86.3**	**84.1**	**94.7**	**90.1**

D.8 GRAD-CAM VISUALIZATION RESULTS ON NIH-CHESTX-RAY14

In this section, we present Grad-CAM visualization results on the NIH ChestX-ray14 dataset, which includes various disease labels such as Pneumothorax, Atelectasis, Mass, Cardiomegaly, Pneumonia, and Effusion. The dataset provides bounding box annotations for certain disease labels, which we use in our evaluations to assess the accuracy of localization. We visualize the regions in the input images that are responsible for the model's predictions on the ground-truth disease labels, comparing the performance of IDS against several baseline models, including MAE [Xiao et al., 2023], OTR [Guo et al., 2022], and REVAR [Jain et al., 2024]. The visualizations in Figure 6 demonstrate that IDS tends to focus more accurately

Table 9: Performance comparisons between IDS models and SOTA baselines on COVIDx (in accuracy). DN represents DenseNet.

Method	Backbone	Covid-19 Sensitivity	Accuracy
COVIDNet-CXR Small [Wang et al., 2020]	-	87.1	92.6
COVIDNet-CXR Large [Wang et al., 2020]	-	96.8	94.4
MoCo v2 [Xiao et al., 2023]	DN121	94.5	94.0
MAE [Xiao et al., 2023]	DN121	97.0	93.5
MAE [Xiao et al., 2023]	ViT-S/16	94.5	95.2
MAE with Synthetic Data		98.0	95.4
MW-Net [Shu et al., 2019]		98.1	96.0
OTR [Guo et al., 2022]		98.0	<u>96.2</u>
IE [Chhabra et al., 2024]		98.0	96.0
CBF [He et al., 2023a]		<u>98.4</u>	96.1
REVAR [Jain et al., 2024]		98.2	<u>96.2</u>
IDS (Ours)		**98.8**	**97.1**
MAE [Xiao et al., 2023]	ViT-B/16	95.5	95.3
MAE with Synthetic Data		98.0	95.5
MW-Net [Shu et al., 2019]		<u>98.5</u>	96.1
OTR [Guo et al., 2022]		98.0	96.1
IE [Chhabra et al., 2024]		98.0	96.0
CBF [He et al., 2023a]		98.1	96.2
REVAR [Jain et al., 2024]		98.2	<u>96.3</u>
IDS (Ours)		**99.0**	**97.3**

on areas inside the bounding boxes provided by the NIH ChestX-ray14 dataset, which correspond to the labeled disease regions. In contrast, the baseline models often activate regions outside the bounding boxes or irrelevant background areas, indicating less precise localization.

Table 10: Performance comparison of various state-of-the-art (SOTA) CNN-based and Transformer-based methods on NIH ChestX-ray14. RN, DN, and SwinT represent ResNet, DenseNet, and Swin Transformer.

Method	Backbone	Pre-training	mAUC
Wang et al.[Wang et al., 2017]	RN50	ImageNet-1K	74.5
Li et al.[Li et al., 2018]	RN50		75.5
LSE-LBA[Yao et al., 2018]	RN&DN		76.1
Thorax-Net[Wang et al., 2019]	R152		78.8
MA[Ma et al., 2019]	R101		79.4
AGCL[Tang et al., 2018]	RN50		80.3
Baltruschat et al.[Baltruschat et al., 2019]	RN50		80.6
DNetLoc [Guendel et al., 2018]	DN121		80.7
CRAL[Guan and Huang, 2018]	DN121		81.6
Seyyed et al.[Seyyed-Kalantari et al., 2020]	DN121		81.2
CAN[Ma et al., 2020]	DN121($\times 2$)		81.7
Hermoza et al.[Hermoza et al., 2020]	DN121		82.1
XProtoNet[Kim et al., 2021]	DN121		82.2
DiRA[Haghighi et al., 2022]	DN121		81.7
ACPL [Liu et al., 2022]	DN121		81.8
SwinCheX [Taslimi et al., 2022]	SwinT		81.0
Categorization [Xiao et al., 2023]	RN50		81.8
Categorization [Xiao et al., 2023]	DN121		82.0
MoCo v2 [Xiao et al., 2023]	DN121	X-rays (0.3M)	80.6
MAE [Xiao et al., 2023]	DN121		81.2
MAE [Xiao et al., 2023]	ViT-S/16	X-rays (0.3M)	82.3
MAE with Synthetic Data			81.8
MW-Net [Shu et al., 2019]			82.0
OTR [Guo et al., 2022]			82.0
IE [Chhabra et al., 2024]			82.1
CBF [He et al., 2023a]			82.1
REVAR [Jain et al., 2024]			82.1
IDS (Ours)			82.7
MAE [Xiao et al., 2023]	ViT-B/16	X-rays (0.5M)	<u>83.0</u>
MAE with Synthetic Data			82.1
MW-Net [Shu et al., 2019]			82.3
OTR [Guo et al., 2022]			82.3
IE [Chhabra et al., 2024]			82.5
CBF [He et al., 2023a]			82.5
REVAR [Jain et al., 2024]			82.5
IDS (Ours)			**83.4**

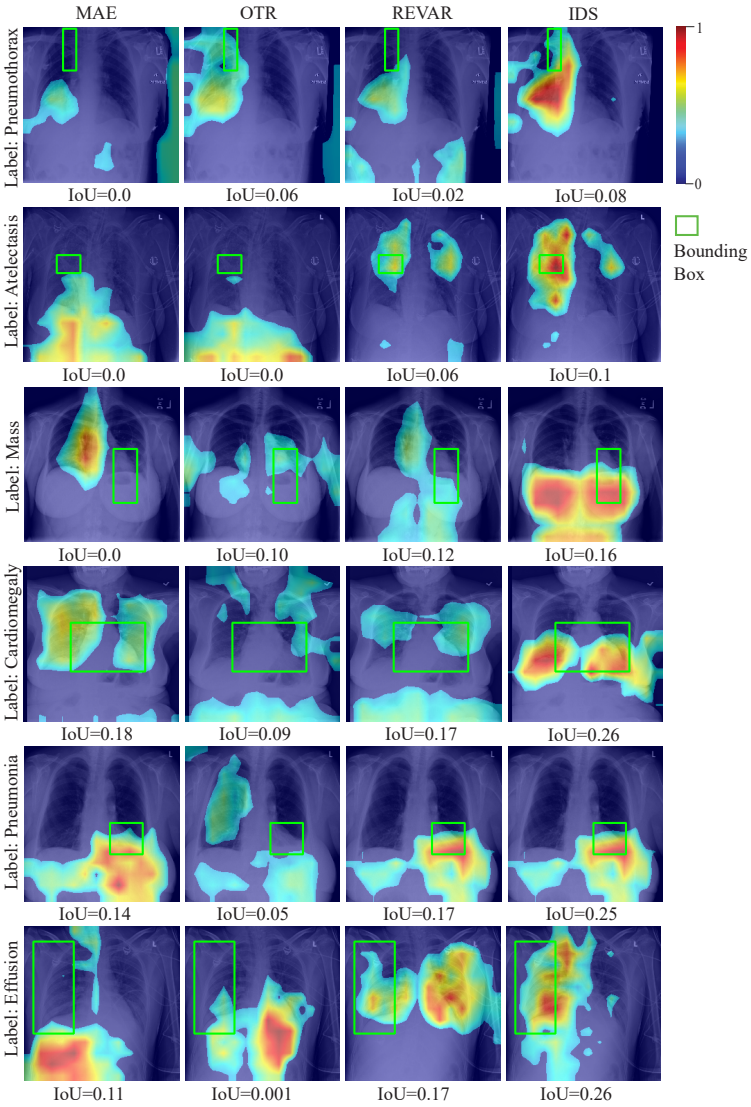


Figure 6: Grad-CAM visualization results on NIH-ChestX-ray14 dataset for various disease labels including Pneumothorax, Atelectasis, Mass, Cardiomegaly, Pneumonia, and Effusion. The visualizations from MAE [Xiao et al., 2023], OTR [Guo et al., 2022], REVAR [Jain et al., 2024], and IDS are shown in the first, second, third, and fourth columns, respectively. The green bounding boxes represent the ground truth regions of interest for each label, and the corresponding IoU score is shown below each image, which quantifies the overlap between the Grad-CAM heatmap and the ground truth bounding box. For each Grad-CAM visualization, higher IoU scores indicate a better localization of the activated regions in relation to the ground truth.