Tackling Distribution Shift - Detection and Mitigation

Laya Rafiee Sevyeri

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy (Computer Science) at

Concordia University

Montréal, Québec, Canada

March 2023

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Laya Rafiee Sevyeri**

Entitled: **Tackling Distribution Shift - Detection and Mitigation**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair
Dr. John Xiupu Zhang

______________________________ External Examiner
Prof. Graham Taylor

______________________________ Examiner
Dr. Marta Kersten

______________________________ Examiner
Prof. Sudhir Mudur

______________________________ Examiner
Prof. Maria Aishy Amer

______________________________ Supervisor
Prof. Thomas Fevens

Approved by ______________________________
Prof. Lata Narayanan, Chair
Department of Computer Science and Software Engineering

February 21st, 2023 ______________________________
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Tackling Distribution Shift - Detection and Mitigation

Laya Rafiee Sevyeri, Ph.D.

Concordia University, 2023

One of the biggest challenges of employing supervised deep learning approaches is their inability to perform as well beyond standardized datasets in real-world applications. Therefore, abrupt changes in the form of an outlier or overall changes in data distribution after model deployment result in a performance drop. Owing to these changes that induce distributional shifts, we propose two methodologies; the first is the detection of these shifts, and the second is adapting the model to overcome the low predictive performance due to these shifts. The former usually refers to anomaly detection, the process of finding patterns in the data that do not resemble the expected behavior. Understanding the behavior of data by capturing their distribution might help us to find those rare and uncommon samples without the need for annotated data. In this thesis, we exploit the ability of generative adversarial networks (GANs) in capturing the latent representation to design a model that differentiates the expected behavior from deviated samples. Furthermore, we integrate self-supervision into generative adversarial networks to improve the predictive performance of our proposed anomaly detection model. In addition, to shift detection, we propose an ensemble approach to adapt a model under varied distributional shifts using domain adaptation. In summary, this thesis focuses on detecting shifts under the umbrella of anomaly detection as well as mitigating the effect of several distributional shifts by adapting deep learning models using a Bayesian and information theory approach.

Acknowledgments

I would like to take this opportunity to wholeheartedly thank Prof. Thomas Fevens, my supervisor, who has been a great support and extremely patient with me throughout my Ph.D. journey. He taught me that research has its ups and downs, and what matters is learning from them.

I would also like to thank my thesis committee, Dr. Marta Kersten, Prof. Sudhir Mudur, Prof. Maria Amer, and my external committee Prof. Graham Taylor for their constructive feedback and their continued support.

I am grateful that I had the opportunity to work with an amazing team of researchers at Imagia during the last year of my Ph.D., which added a new dimension to my academic career. I want to thank Mohammad Havaei for his constant support and outstanding mentorship during my internship.

My deepest gratitude goes to my partner Farhood Farahnak, my first mentor who introduced me to deep learning and has always been there for me since the beginning of my journey. My parents, Maman and Baba, who always support me and motivate me to pursue my dreams. My sisters Leili and Sara who, despite the distance, are always my best friends, and without their unconditional support and care, I couldn't reach my goals.

In the last six months, the women and men in my homeland Iran have been fighting for freedom. They were all on my mind while I was writing my thesis, and their courage was my biggest inspiration.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Access to large-scale datasets is the pillar of the rapid growth of deep neural networks. In this regard, the necessity of proper data annotation limits supervised learning and makes unsupervised and self-supervised learning more in-demand. On the other hand, discriminative deep learning models are predictive models only limited to a decision boundary. Despite their success in many practical applications, uncovering the distribution of data is beyond the purpose of these models. In addition, inevitable discrepancies between training and test data in real-world applications, known as distribution shifts, hurt the generalizability of deep learning models. Performance drop and lower generalizability are even more drastic in applications in medical imaging, where data acquisition and annotation are often very expensive. The disparity between training and test sets' distribution might be either a sudden and unusual change in the distribution associated with few examples (out-of-distribution also known as OOD) or a permanent change in the distribution visible in the entire test set. Hence, the path to address each of these types of shifts can be defined differently.

The issues mentioned above are the inspirations for creating fundamental tools for machine learning systems. Unsupervised learning and, in particular, generative models open new doors to discovering the underlying structure of the data, i.e., data distribution. While unsupervised learning mitigates the problem of costly annotated data, generative models, that are mostly unsupervised, focus on sample generation via the learned distribution (Goodfellow et al., 2014; Kingma, Salimans,

& Welling, 2015; Salakhutdinov & Hinton, 2009).

The problem of identifying changes in parts of the test set defines anomaly detection (AD), i.e., finding those samples which do not fit the training data distribution (Chandola, Banerjee, & Kumar, 2007). While the idea of anomaly detection spreads to different research areas, arising generative models help in learning the distribution of normal data; due to limited labeled data and insufficient knowledge of known anomalies in disease detection, unsupervised anomaly detection becomes an exciting tool in medical imaging (Schlegl, Seeböck, Waldstein, Schmidt-Erfurth, & Langs, 2017).

Chapter 4 and 5 of this thesis will focus on detecting anomalies in different and challenging domains. Particularly, as one of our studies, we want to investigate the feasibility of anomaly detection on small medical datasets, which are relatively common in medical imaging. In addition to a medical dataset, we experiment with more complex and larger datasets. Among various generative models, we explore the benefit and difficulties of the generative adversarial network (GAN) (Goodfellow et al., 2014) on anomaly detection. Although GANs have proven to be effective, several obstacles make their training difficult and, therefore, lower performance compared with non-generative state-of-the-art approaches. Similar to other deep neural networks, generative models also suffer from catastrophic forgetting, i.e., a problem when the model forgets about previous tasks/classes when dealing with a new one. Furthermore, we conduct an investigation into possible remedies for these models' limitations, mainly catastrophic forgetting and mode collapse, and introduce a generative model that benefits from self-supervision to address these issues.

Even though identifying shifts in the test distribution and pointing out those samples that show signs of divergence from training distribution is important, anomaly detection models still need strategies to further indicate where shifts originate and how to adapt the model under the presence of shifts. Besides, changes between the training and test distributions can easily happen. For instance, a minor alteration in the distribution of labels between two domains, i.e., a change in the proportion of a single class between source and target domains in a classification task, may decrease the performance and reduce the generalizability. Considering $p(X, Y)$ as the joint distribution of a domain, where X and Y define the input and the output respectively, any changes in the input, the output, or both in the target domain introduce shift. Consequently, domain adaptation was introduced as a tool to help the models adapt themselves to various shifts.

In Chapter 6 of this thesis, we focus on mitigating the distribution shifts rather than detecting them. Specifically, we try to mitigate two major sources of shift: covariate shift (drift in the input) and label distribution shift. This study inspects multiple medical, natural, and synthetic datasets under distributional shifts. Unlike many domain adaptation approaches, our solution addresses shifts while mitigating privacy and storage concerns typical to domain adaptation.

1.2 Distribution Shifts

Generally, machine learning (ML) models, under the i.i.d. assumption that states random variables are independent and identically distributed, rely on a large set of examples drawn from the same distribution to solve a learning task. Given a set of labeled examples $(x, y) \in X, Y$ drawn from the training distribution p and $(x', y') \in X', Y'$ from the test distribution q, the i.i.d. assumption implies p and q to have the same distribution (stationary environment), i.e., $p = q$. The training and test sets can be referred to as source and target distributions in this definition.

While under this assumption, ML models flourish and unprecedented breakthroughs happened in various fields, the disparity in the source and target distribution is inevitable. Difference between distributions, refer to as distribution shift, degrades the performance and hurt the generalizability of ML models. The joint distribution between data points and their corresponding labels of the two domains varies by different sources. In the following, we introduce trivial shifts generally known as distribution shifts (Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2008).

Covariate Shift

From the statistical point of view, the joint distribution $p(x, y)$ can be decomposed in $p(x, y) = p(x)p(y \mid x)$, where $p(x)$ defines marginal distribution and $p(y \mid x)$ defines the conditional label distribution.

Covariate shift refers to the changes in the covariate x between source and target, $p(x) \neq q(x)$, while the conditional distribution remains fixed, i.e., $p(y \mid x) = q(y \mid x)$. Therefore, the joint distribution of source and target changes, i.e., $p(x, y) \neq q(x, y)$.

Label Distribution Shift

The joint distribution also defines as $p(x, y) = p(y)p(x \mid y)$. *Label distribution shift* which is also called as *target shift* or *prior probability shift* indicates changes over label y between source and target domains, $p(y) \neq q(y)$, while the conditional covariate distribution is invariant, $p(x \mid y) = q(x \mid y)$. Similarly, this change leads to a shift in the joint distribution of the two domains.

Concept Drfit

Concept drift refers to shifts where the changes stem from conditional distributions rather than marginals. Particularly when either conditional distribution over covariate or label changes, i.e., $p(x \mid y) \neq q(x \mid y)$ or $p(y \mid x) \neq q(y \mid x)$, while the marginals remain fixed, i.e., $p(x) = q(x)$ and $p(y) = q(y)$, *concept drift* happens.

In addition to the aforementioned shifts, there are other types of shifts that are less investigated in the literature. Before explaining the necessity of detecting shifts and different approaches for them, we briefly introduce the primary reasons for the emergence of shifts.

The two common causes of shifts between the source and target distribution are i) sample selection bias and ii) non-stationary environments. The former is related to the discrepancy in the distribution, which is due to the data acquisition and annotation biases. The biases might cause misrepresentation of the environment. The latter originates from real-world non-stationary environments. This feature is associated with the fact that real world phenomena continuously vary from both temporal and spatial perspectives.

1.2.1 Anomaly Detection

The process of finding patterns in the data which do NOT resemble the expected behavior is known as anomaly detection (AD). Particularly, the problem of detecting distribution shift can be reduced to whether a sample comes from the same distribution as the training data, which has been widely studied under the name of anomaly detection in the literature (Chandola, Banerjee, & Kumar, 2009). Despite the fact that anomaly detection approaches are only limited to identifying shifts between two domains rather than detecting specific types of shifts, they are well-established for different machine learning tasks.

Anomaly detection, which is often also referred to as out-of-distribution (OOD) detection, can explore two main approaches; novelty or outlier detection. Even though the names frequently have been used interchangeably in the literature, there is a narrow line separating the two. Novelty detection mostly involves a detection model where the training dataset only contains a particular distribution, and no abnormalities are present. In comparison, outlier detection denotes a learning model where the training dataset may include abnormalities.

Anomaly detection approaches have evolved over time from classical approaches such as one-class support vector machines (Schölkopf et al., 1999), density estimation (Breunig, Kriegel, Ng, & Sander, 2000), and isolation forest (F. T. Liu, Ting, & Zhou, 2008), to deep learning based approaches (Golan & El-Yaniv, 2018) and GANs (Schlegl et al., 2017).

1.2.2 Transfer Learning

As stated earlier, one of the primary keys to the advancement of ML models is access to abundant annotated data, which is sometimes very expensive and time-consuming. These models often assume the training and test samples to be drawn from the same distribution, also known as the i.i.d. assumption. Despite the unprecedented success of machine learning in many practical applications, these limitations become obstacles in real-world applications and deteriorate the performance of ML models. To fill the performance gap between training and test, transfer learning (TL) approaches are proposed.

Given a source domain $\mathcal{D}_s$ and learning task T_S, a target domain $\mathcal{D}_t$ and learning task T_T, transfer learning aims to improve the learning of the target predictive function $f_t(.)$ in $\mathcal{D}_t$ using the knowledge in $\mathcal{D}_s$ and T_S, where $\mathcal{D}_s \neq \mathcal{D}_t$, or $T_S \neq T_T$ (Pan & Yang, 2009).

Considering the learning task as learning a joint distribution $P(x, y)$, then transfer learning can be defined as a process of adapting a model trained on one joint distribution to another joint distribution. Explicitly, in transfer learning a model that is developed for a task will be reused as a starting point for a second task. However, in domain adaptation which is a branch of transfer learning, retraining the model might be necessary for the adaptation.

Various types of transfer learning are defined based on the similarity of the task between source and target domains and the availability of labeled data in the target domain (Pan & Yang, 2009). *Inductive transfer learning* refers to a learning method where the task between source and target domains are different. This category includes previous work based on reweighting strategy, also known as instance-based transfer learning (Jiang & Zhai, 2007; Sugiyama, Nakajima, Kashima, Buenau, & Kawanabe, 2007). Whereas in *transductive transfer learning*, the task remains fixed while the domains are different like (Argyriou, Evgeniou, & Pontil, 2006; Raina, Battle, Lee, Packer, & Ng, 2007). *Unsupervised transfer learning* is similar to the inductive approach in the sense that the target task varies from the source, but unlike inductive transfer learning, which has access to a set of labeled samples in the target domain, no labeled data is available in unsupervised transfer learning. Similarly, transductive transfer learning only has access to unlabeled target data. From the definition, domain adaptation is a type of transductive transfer learning.

1.3 Problem statement

This research study focuses on distribution shifts and investigates novel approaches for detecting and mitigating them.

Anomaly detection has been explored broadly in various areas of machine learning (see Chapter 3). Similar to distribution shift detection, anomaly detection models investigate and reveal deviations from the source domain (training set). Such an approach can be utilized to train models that are able to distinguish one class among the others (cats versus other pets) as well as a tool to capture unknown anomalies, e.g., within the process of disease detection, benefiting the lack of known anomalies and limited labeled data.

Anomaly detection approaches, however, are bounded only to detecting shifts. Hence further information concerning the types of shifts and approaches to mitigate them requires other techniques.

In many applications, distribution shift detection is not a priority. In these cases, we prefer an existing model that can maintain its performance under possible shifts. Therefore, transfer learning techniques should be used to address these problems.

1.4 Research Objectives

The problem of limited labeled data, limited known anomalies, and unbalanced data are more severe in medical imaging. As a result, a model that built upon these limitations might perform poorly given possible shifts. We conduct research on using anomaly detection-based approaches for shift detection. We deploy adversarial training to propose a simple model to detect anomalies in multiple domains. We further examine the negative impact of mode collapse and catastrophic forgetting in our adversarial training and investigate the effect of integrating self-supervised learning in our proposed adversarial model to improve the performance of our detection model.

As stated earlier, in some cases, mitigating the effect of distribution shifts might have a higher priority than detecting them. In our last research, we focus on deploying a model that can perform under different distributional shifts.

Our main research questions are as follows:

(1) How can an adversarial training approach help us detect shift while overcoming the problem of limited labeled data and unknown anomalies?

(2) How does integrating self-supervised learning into adversarial learning help to improve performance while mitigating mode collapse and catastrophic forgetting of our generative adversarial network-based model?

(3) How to propose a model that maintains its performance under covariate and label distribution shifts?

(4) How does ensemble training improve the performance of a domain adaptation model under different distributional shifts when the source data is not available during the adaptation phase?

Based on these research questions, the major parts of our research focus on identifying deviations in the target/test data where shift detection transforms into the problem of detecting anomalies without any labeled data. On the other hand, our second research direction involves transfer learning to design a model to perform under various shifts. Our answer to the first question is an unsupervised model targeted at identifying samples that are not coming from the training distribution in the context of natural and medical images. This research evaluates different scoring functions to distinguish normal samples from abnormal samples. In the next study and to answer the second question, we introduce a similar adversarial model which benefits from self-supervised learning to address the two major limitations of GANs. This research focuses on different natural image benchmarks. In the last research study, we address the last two questions simultaneously. We propose an ensemble model with more diverse members and a weighted regularization approach in a source-free domain adaptation model to mitigate the effect of covariate and label distribution shifts. Besides the natural and synthetic images, our solution also evaluates on a medical dataset under two major distribution shifts.

1.5 Contributions

In this research study, we focus on applying new methods in detecting and mitigating different types of distributional shifts in various domains.

The key advantage of using deep models over the classical machine learning approaches is their ability to grasp the hidden information that lies in the data without further feature engineering. This property is obtained mainly from very large-scale data. In supervised learning tasks, having a large amount of data with their corresponding ground truth labels is not always accessible, especially in medical imaging, where data acquisition and annotation are often costly and time-consuming. Given the limitations of supervised learning, considering unsupervised training seems promising.

In all of our research studies, given an unsupervised setting, we explore the difficulty and possibilities of different deep learning approaches in the presence of various distributional shifts.

- The first study proposes a new unsupervised deep generative model accompanied by a new anomaly score function to identify anomalies in the context of images. To further investigate the effectiveness of our model, we evaluate it on two benchmark datasets as well as a public medical dataset. The proposed model achieved the highest performance among existing approaches. We show experimentally that our proposed model can perform well even on a small-sized dataset which is not very appealing in many deep learning tasks. The result of this study was published at 29th International Conference on Artificial Neural Networks (ICANN2020).
  - ○ This work was also presented as a poster at Montréal AI Symposium (MAIS) in September 2020.

- The second study introduces a new unsupervised anomaly detection model inspired by contrastive learning to mitigate the two common issues of generative adversarial networks. Our model simultaneously addresses mode collapse and catastrophic forgetting and significantly improves the performance of GAN-based anomaly detection models. We conduct various experiments on multiple benchmark datasets with different dataset sizes. This research was published at the 21st International Conference of Image Analysis and Processing (ICIAP2021) and received an ”NVIDIA Winner Prize”.

  ◦ Accepted as a poster in Bayesian Deep Learning NeurIPS workshop (NeurIPS 2021)

- The last research introduces an effective domain adaptation model based on ensemble learning with a weighted regularization scheme in the presence of covariate and label distribution shifts. In this setting, unlike most domain adaptation approaches, access to the source data is only limited to a model induced from them. Extensive experiments on multiple domains of natural, synthetic, and medical demonstrate the effectiveness of our model. This work is currently under review in the Machine Learning journal.

List of other contributions as a coauthor in chronological order:

- The Concordia NLG Surface Realizer

  ◦ Farahnak, F., Rafiee, L., Kosseim, L., Fevens, T. ”The Concordia NLG Surface Realizer at SRST 2019.” Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019). 2019.

- Surface Realization Using Pretrained Language Models

  ◦ Farahnak, F., Rafiee, L., Kosseim, L., Fevens, T. ”Surface realization using pretrained language models.” Proceedings of the Third Workshop on Multilingual Surface Realisation. 2020.

- AdaBest: Minimizing Client Drift in Federated Learning via Adaptive Bias Estimation

  - Varno, F., Saghayi, M., Rafiee Sevyeri, L., Gupta, S., Matwin, S., Havaei, M. (2022). ”AdaBest: Minimizing Client Drift in Federated Learning via Adaptive Bias Estimation.” European Conference on Computer Vision. Springer, Cham, 2022.

- Learning from Uncertain Concepts via Test Time Interventions

  - Sheth, I., Abdul Rahman, A. , Rafiee Sevyeri, L., Havaei, M., Ebrahimi Kahou, S. ”Learning from uncertain concepts via test time interventions.” Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022

Chapter 2

Background

In this Chapter and to follow this research study, some preliminary definitions of machine learning (ML) and deep learning (DL), as well as several well-known deep generative models, will be reviewed.

2.1 Supervised Learning versus Unsupervised Learning

A machine learning algorithm refers to the types of algorithms that are able to learn from their experience. One of the best definitions of learning is proposed by Mitchell (1997):

> ”A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task T, which is measured by P, improves with experiences E.”

The learning process can be done in different ways. Considering the data and types of information it provides, three major learning paradigms can be defined: i) supervised, ii) unsupervised, and iii) semi-supervised learning.

Supervised learning is a class of machine learning tasks where a model trained on a training dataset consists of a set of samples ($x_i \in X$) with their corresponding ground-truth labels ($y_i \in Y$), and the task is to learn a mapping function $y_i = f(x_i)$ from the input to the output. The goal is to approximate the mapping function with minimum error, therefore given a new input data (x_j), the model can predict the corresponding output variables (y_j) using the mapping function (Bishop, 2006).

It is called supervised learning because the whole process of an ML algorithm which is learning from the training dataset, can be thought of as a teacher supervising the learning process. In this kind of learning procedure, we have the correct answers to the training data, so the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Supervised learning problems are further grouped into regression and classification problems.

- Classification: The output variable in a classification problem is a category, such as "red" or "blue" or "disease" and "no disease".
- Regression: The output variable in a regression problem is a real value, such as "dollars" or "weight".

On the other hand, an unsupervised learning task only has access to the input variables, and their corresponding ground-truth are not available. The goal of unsupervised learning is to model the underlying structure or distribution of the data in order to learn more about the data. Unlike supervised learning, there is no supervision for the given training data. Unsupervised learning problems can be further grouped into clustering and association problems (Bishop, 2006).

- Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by their purchasing behavior.
- Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy x also tend to buy z.

However, the two categories mentioned earlier are rather old, and the latest and more advanced approaches for unsupervised learning do not fit into them. Dimensionality reduction models such as Principal component analysis (also known as PCA) (Pearson, 1901; Wold, Esbensen, & Geladi, 1987) and generative models like autoencoder (Ballard, 1987) and generative adversarial networks (Goodfellow et al., 2014) are as such unsupervised approaches beyond the category of clustering and association.

Rather than supervised and unsupervised learning, there is an intermediate level of learning which is called semi-supervised learning, where the ground-truth labels are available only for usually small portions of the dataset.

Recently, a new sub-category was added to the existing learning paradigms called self-supervised learning. In the self-supervised learning paradigm, often an auxiliary task, also known as surrogate or pretext, defines where labels are readily extractable from the data without any human intervention. The strong supervision signals within the surrogate tasks enable the model to leverage the objective function similar to the way it is done in supervised learning.

2.2 Deep Learning

Machine learning technology powers many aspects of modern society, from web searches and content filtering on social networks (Chau & Chen, 2008; Vanetti, Binaghi, Carminati, Carullo, & Ferrari, 2010) to recommendations on e-commerce websites (Zhao, Zhang, Friedman, & Tan, 2015), and it is increasingly present in consumer products such as cameras and smartphones. Machine learning systems are used to identify objects in images (Dalal & Triggs, 2005; Lowe, 1999; Viola & Jones, 2001), transcribe speech into text (Ganapathiraju, Hamaker, & Picone, 2000; Woodland & Povey, 2002), match news items, posts or products with users' interests, and select relevant results of search (Joachims, 2002; Mohan, Chen, & Weinberger, 2011). Increasingly, these applications often use a class of techniques called deep learning.

Conventional machine learning techniques are limited in their ability to process natural data in their raw form. For decades, constructing a pattern recognition or machine learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning refers to a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep learning methods are accounted as representation learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned (Goodfellow, Bengio, & Courville, 2016).

The emergence of deep learning models was back to the introduction of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a convolutional neural network (CNN) (LeCun, Haffner, Bottou, & Bengio, 1999) model and the winner of ImageNet challenge (J. Deng et al., 2009) in 2012. After 2012, the field witnessed a handful of research on discovering the ability of deep models in speech recognition (Chan, Jaitly, Le, & Vinyals, 2016), visual object recognition (He, Zhang, Ren, & Sun, 2015), object detection (Girshick, Donahue, Darrell, & Malik, 2014; Redmon, Divvala, Girshick, & Farhadi, 2016), and many other domains such as drug discovery (H. Chen, Engkvist, Wang, Olivecrona, & Blaschke, 2018) and genomics (Park & Kellis, 2015; Quang & Xie, 2016) ended to dramatically improved the state-of-the-art. With the help of the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986), neural networks are able to compute the gradient and, therefore deep learning models to uncover complex structures in large datasets. Optimization algorithms like stochastic gradient descent (SGD) use the gradient to facilitate changes in a neural network's internal parameters to increase the predictive model accuracy by decreasing a surrogate loss function.

2.3 Generative versus Discriminative Models

There are two main approaches to doing a machine learning task from the statistical machine learning point of view: **generative** and **discriminative** approaches.

Assuming an input variable x and its corresponding output y from input space $\mathcal{X}$ and output space $\mathcal{Y}$ respectively, a generative approach attempts to learn the joint probability distribution $p(x, y)$ (Ng & Jordan, 2002). A more common definition would be a model which describes how data is generated in terms of a probabilistic model. Whereas the discriminative approach tries to find the conditional probability of the target variable y given an input x; $p(y \mid x = x)$. Discriminative models, also known as conditional models, are a class of models used in machine learning for modeling the dependence of unobserved variables, i.e. target, on observed variables, i.e. input.

Within a probabilistic framework, this is done by modeling the conditional probability distribution. If the primary goal is prediction, then discriminative models, which directly estimate $p(y|x)$, are found to be empirically superior because they attack the problem directly. However, discriminative models tend to gain little understanding of the data.

Discriminative models do not allow one to generate samples from the joint distribution of observed and target variables, i.e., $p(x, y)$, as opposed to generative models. However, for predictive tasks such as classification and regression that do not require the joint distribution, discriminative models can yield superior performance because they have fewer variables to compute.

The most compelling successes of deep learning came from the discriminative models, where they map a high dimensional input into a target label (Dosovitskiy et al., 2021; He, Zhang, Ren, & Sun, 2016; Tan, Pang, & Le, 2020). On the other hand, generative models that have to deal with the data distribution didn't gain much success until recently (Goodfellow et al., 2014; J. Ho, Jain, & Abbeel, 2020; Keller & Welling, 2021; Vaswani et al., 2017; Zhu, Park, Isola, & Efros, 2017).

In generative models that utilize joint distribution, they can either directly use the joint distribution of the data, i.e., $p(x, y)$, or follow the Bayes theorem (Eq. 1):

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(X)} \tag{1}$$

where C_i is the label of the i-th class, x is the input, $p(C_i)$ is the prior which is the probability of class C_i, and $p(x|C_i)$ is the likelihood of input x belongs to class C_i. By having the data distribution, we are able to generate synthetic data in the input space. Due to the difficulty of approximating the joint distribution, there was not much progress in generative-based approaches as we witnessed in the discriminative approaches until earlier (Goodfellow et al., 2014; Kingma & Welling, 2013).

However, we have to notice that even if the model knows the true probability distribution that generates the data, it may still incur some errors on some inputs. This happens due to the inherent noise in the data distribution which is referred to as Bayes error in the literature (Fukunaga, 1990).

Unsupervised learning is a very broad term that encompasses many different ways of finding structure in unlabeled data. Generative modeling means building a model that can generate new examples that come from the same distribution as the training data or look at an input example and report the likelihood of it being generated by that distribution. This means generative modeling is a kind of unsupervised learning alongside other kinds of unsupervised learning like clustering and dimensionality reduction approaches.

A generative model can be designed to generate images, text, and voice. Parts of this research study will mainly focus on generative models in the context of images. Some of the more common generative approaches which have been broadly used in the literature in each context will be reviewed in chapter 3.

2.4 Deep Generative Models

Advancements in generative models following the emergence of deep learning rapidly increased. In this section, we briefly introduce a few famous deep generative models that have been the focus of our research. Since reviewing ongoing research on generative models is beyond the scope of this research, we encourage the readers to read more about recent works in literature, e.g., Diffusion models (J. Ho et al., 2020) and GFlowNets (Bengio, Jain, Korablyov, Precup, & Bengio, 2021).

Autoencoder (Ballard, 1987) is an unsupervised learning technique that leverages artificial neural network (ANN) to learn a representation for a set of data where the task can be dimensionality reduction, feature learning, and recently learning generative models of data.

The simplest form of an autoencoder is a feed-forward neural network similar to a multilayer perceptron (MLP) but considering the fact that the output layer should have the same number of nodes as the input layer to reconstruct its own inputs (instead of predicting the target value y given inputs x).

An autoencoder consists of a hidden layer h, often called the bottleneck. A bottleneck forces the model to learn compressed representations of the input data by limiting the amount of information that can pass over the network. The network may be viewed as two different components: an encoder function $h = f(x)$ to map the input to a condensed representation (encoded) and a decoder function $r = g(h)$ that generate a reconstructed sample from the encoded representation (see Fig. 2.1). These kinds of networks are usually restricted in ways that only allow them to copy the features that resemble the training data, pushing towards learning more useful properties of the data (Goodfellow et al., 2016).

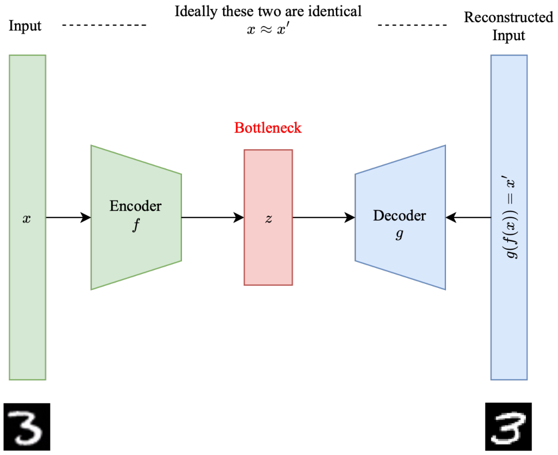


Figure 2.1: Autoencoder architecture; given a sample x, autoencoder tries to learn an encoded representation to reconstruct the input.

In order to train these types of generative models, a loss function defined as $L(x, g(f(x)))$ will be minimized throughout learning. The loss function is known as reconstruction error which measures the differences between the original input and the resulting reconstructed output. The reconstruction error function varies based on the type of input data. In the case of binary inputs, the loss function can be the cross-entropy loss:

$$L(x, g(f(x))) = -\sum_{i=1}^{n}(x_i \ \log(g(f(x_i))) + (1 - x_i) \ \log(1 - g(f(x_i)))) \tag{2}$$

or it can be defined as mean squared error (MSE) in the case of real-valued input:

$$L(x, g(f(x))) = \frac{1}{n}\sum_{i=1}^{n}(x_i - g(f(x_i)))^2 \tag{3}$$

Rather than limiting the dimension of the bottleneck, regularized autoencoders use a regularized loss function that encourages the model to have other properties besides the ability to copy its input to the output. Sparse autoencoders, Denoising autoencoders (Vincent, Larochelle, Bengio, & Manzagol, 2008), and Contractive autoencoders (Rifai, Vincent, Muller, Glorot, & Bengio, 2011) are families of regularized autoencoders.

Among several variations of autoencoder, Variational Autoencoder (VAE) (Kingma & Welling,

2013) is a probabilistic model that, unlike the vanilla autoencoder, generates a probability distribution over the latent attributes.

Generative Adversarial network (GAN) (Goodfellow et al., 2014) is a two-player minimax game defined by its two competing models; a generator G and a discriminator D. In this framework, two models are trained simultaneously. The discriminator is trained to discriminate between the samples coming from true data distribution and the generator distribution. In contrast, the generator tries to decrease the probability of being fake by approximating the data distribution from a much simpler distribution, e.g., a Gaussian or Uniform distribution. That is to say, during the training, the generator trains in a way to lead the discriminator to make more mistakes. Whereas the discriminator trains to maximize the probability of predicting the true labels, where the labels are either real or fake.

The competition between these two adversaries drives them to improve their methods. By definition, the ideal stopping point is when G captures the distribution of the training data, and D can not distinguish between the generated data and training data anymore, i.e., returning the probability of $\frac{1}{2}$ for each sample.

Even though there is no limitation on the types of the generator or the discriminator models, Goodfellow et al. (2014) suggested that both G and D be multilayer perceptrons so they can be trained using the backpropagation technique.

The primary purpose of GAN is to mimic the distribution of training data and, therefore, the ability to generate samples drawn from the learned distribution. Samples are generated from a random noise drawn from simpler distributions than the training data, typically a Gaussian or uniform (a schematic view of the architecture of a GAN model is shown in Fig. 2.2).

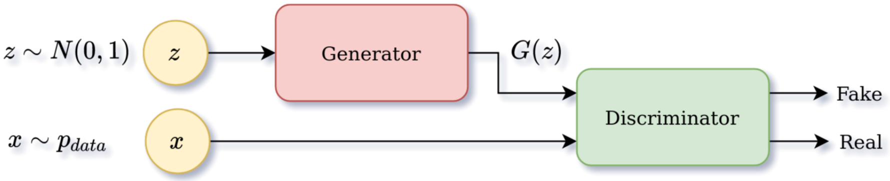


Figure 2.2: GAN architecture, where z is the noise sampled from a Gaussian distribution, $G(z)$ is the generated image from noise z, and x is a training sample drawn from p_{data} distribution.

Given G and D as two neural networks, the training of GANs can be formulated in this way. In this framework, the generator G tries to learn the distribution over data x and a prior on input noise variable, which is defined as $p_z(z)$. $G(z;\theta_g)$ is the mapping function where G is a neural network parameterized by θ_g. On the other side, $D(x;\theta_d)$ defines the mapping function for the discriminator D. Here p_g and p_{data} represent the generated and the training data distribution respectively. The discriminator's result is a single scalar representing the label of the input data, i.e., $D(x)$ defines the probability of x belonging to training data distribution rather than p_g. While D trains to maximize the probability of assigning the true labels, both to the samples that come from the training data or the generated samples, G trains to minimize $\log(1 - D(G(z)))$. Hence, D and G play a minimax game to optimize the following objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

Since a discrimination task is much easier than the generation, this objective function might end up with a poor generator. Goodfellow et al. (2014) suggested to reformulate Eq. 4 by replacing minimizing $\log(1 - D(G(z)))$ with maximizing $\log(D(G(z)))$.

As mentioned earlier, the training procedure of the GAN is very similar to the minimax game, where both competitors try to improve their models to beat each other. From the game theory, the model converges when the discriminator and the generator reach a Nash equilibrium (Osborne & Rubinstein, 1994). In this adversarial training, the global optimum is $p_g = p_{data}$, where the generator can capture the data distribution, and it is achievable if both D and G have enough capacity.

Even though GAN has gained quite a lot of success in many different areas (T. Chen, Zhai, Ritter, Lucic, & Houlsby, 2019; Ledig et al., 2017; Mahapatra, Bozorgtabar, & Garnavi, 2019; Schlegl et al., 2017; H. Zhang et al., 2017), it still suffers from unstable and hard training procedure (Lucic, Kurach, Michalski, Gelly, & Bousquet, 2018), mode collapse (failure of GANs to capture important features of a target distribution) (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017), diminished gradient (meaning that the discriminator gets too successful that the generator's gradient vanishes and learns nothing), catastrophic forgetting (Kemker, McClure, Abitino, Hayes, & Kanan, 2018), and its highly sensible behavior to the choice of hyperparameters (Lucic et al., 2018).

2.5 Calibration and Uncertainty in Neural Network

Assuming a supervised classification task and a given input $X \in \mathcal{X}$, a label $Y \in \mathcal{Y}$, and $\phi(X;\theta) = (\hat{Y}, \hat{P})$ a neural network parameterized by θ, $\hat{Y}$ represent the predicted class by ϕ and $\hat{P}$ defines the confidence associated to the class prediction (C. Guo, Pleiss, Sun, & Weinberger, 2017). In this setting, we wish that $\hat{P}$ to be well-calibrated. In other words, if the estimated confidence represents the true probability, it is then calibrated. For instance, given 100 examples and a predictive model with the confidence of 0.95 for each sample, the expectation is that 95 samples to be correctly classified.

Uncertainty in neural networks is closely related to the confidence of these models. C. Guo et al. (2017) showed that even though current large-scale neural networks like ResNet110 (He et al., 2016) achieve higher accuracy compared with shallow networks, they tend to be overconfident on incorrectly labeled, noisy, or unseen data and therefore less well-calibrated. Uncertainty in neural networks has two types; aleatoric uncertainty and epistemic uncertainty. The former, which is also referred to as data uncertainty, is an inherent property of the data distribution, e.g., noise or perturbation in the input data. On the contrary, epistemic uncertainty, also refers to as model/knowledge uncertainty, occurs due to inadequate knowledge.

Brier Score (BS) (Brier et al., 1950), Expected Calibration Error (ECE), and Maximum Calibration Error (MCE) (Naeini, Cooper, & Hauskrecht, 2015) are standard evaluation metrics to measure the calibration capability of a predictive model. Considering a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, y_n and $p(y \mid x_n, \theta)$ define the ground truth and the predicted probability, Brier score measures the accuracy of predicted probabilities (Eq. 5):

$$BS = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} [p(y \mid \mathbf{x}_n, \theta) - \delta(y - y_n)]^2. \tag{5}$$

Expected Calibration Error (ECE), shown in Eq. 6, measures the difference between predicted probabilities and accuracy.

$$ECE = \sum_{b=1}^{B} \frac{s_b}{N} |acc(b) - conf(b)| = \sum_{b=1}^{B} \frac{s_b}{N} \left| \sum_{n \in b} ((y_n = \hat{y}_n) - (p(\hat{y}_n \mid \mathbf{x}_n, \theta))) \right| \tag{6}$$

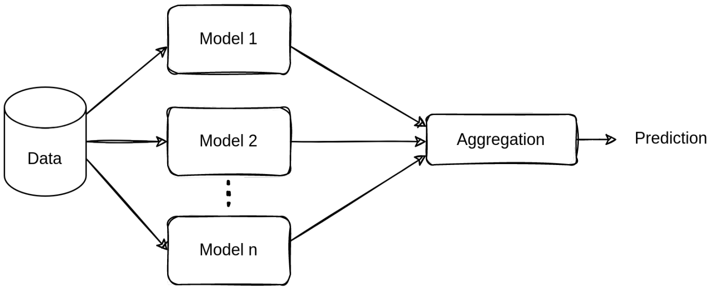


Figure 2.3: An abstract overview of an ensemble approach.

where B defines the number of bins, N total number of samples, s_b indicates the number of samples in bin b, and $\hat{y}_n = \text{argmax}_y p(y \mid \mathbf{x}_n, \theta)$. The accuracy $acc(b) = \sum_{n \in b} \frac{(y_n = \hat{y}_n)}{s_b}$ is also called "observed relative frequency", while the confidence $conf(b) = \sum_{n \in b} \frac{p(\hat{y}_n \mid \mathbf{x}_n, \theta)}{s_b}$ refers to "average predicted frequency". Maximum Calibration Error (MCE) measures the maximum difference between predicted probabilities and accuracy among the bins. In addition to these three metrics, there are several other scores to estimate the calibration of a model.

2.6 Ensemble Learning

Ensemble learning is a learning approach that seeks better predictive performance by combining the predictions from multiple models (Opitz & Maclin, 1999; Rokach, 2010) (see Fig. 2.3). Classical ensemble models include Bagging (Breiman, 1996), Stacking (Wolpert, 1992), and Boosting (Freund, Schapire, et al., 1996). Bagging averages the predictions of multiple decision trees on different samples of the same dataset, while stacking uses another model to learn the best approach to aggregate the ensemble members' predictions. Unlike the other two, boosting improves the predictions of the previous members and outputs a weighted average of the predictions.

Rather than traditional approaches, ensembles have also been introduced in deep learning via deep ensemble (Lakshminarayanan, Pritzel, & Blundell, 2017). Deep ensemble, with the support of experiments and theoretical analyses, showed that ensemble mitigates calibration and uncertainties in neural networks.

2.7 Evaluation Metrics

This section will briefly present evaluation metrics used in this research study.

The effectiveness of a binary classification model usually measures with some defined metrics known in the literature. In a medical imaging task, a binary classification model has two classes of benign, i.e., healthy, and malignant, i.e., unhealthy. The positive outcome defines the unhealthy class, whereas the negative outcome represents the healthy class. Considering these two classes, the true/false positive/negative can be defined this way.

True positive (*tp*) refers to correctly identified malignant samples, and true negative (*tn*) refers to samples that are correctly classified as benign. On the other hand, a false positive (*fp*) denotes samples incorrectly identified as malignant, while a false negative (*fn*) denotes samples incorrectly classified as benign.

Therefore, in order to measure the performance of a binary classification model, the following metrics will be calculated.

- **Sensitivity**, also called recall, measures how well the model detects the unhealthy (malignant) samples; $\frac{tp}{tp+fn}$
- **Specificity** measures how well the model detects the healthy (benign) samples; $\frac{tn}{tn+fp}$
- **Precision** defines the fraction of actual unhealthy (malignant) samples among all of those predicted as unhealthy; $\frac{tp}{tp+fp}$
- **F1-measure** is the harmonic mean of precision and sensitivity
- **Accuracy** measures how well the model detects both classes; $\frac{tp+tn}{tp+tn+fp+fn}$
- **AUC** or the area under the Receiver Operating Characteristic (ROC) curve is a measure of how well a parameter can distinguish between healthy and unhealthy classes, where the diagnostic performance of a test or the accuracy of a test to discriminate unhealthy cases from healthy cases is evaluated using ROC curve analysis.

Chapter 3

Literature Review

3.1 The History of AI and the Deep Learning Emergence

The emergence of artificial neural networks (ANN) dates back to the introduction of the perceptron in 1958 (Rosenblatt, 1958). Upon the emergence of perceptrons, many believed they could solve all the problems until 1969, where Minsky and Papert (1969) proved only linearly separable functions could be presented by perceptron, and they can not even solve a simple XOR problem. The research halted for almost 20 years until when the multilayer perceptron with nonlinear activation function and backpropagation algorithm (Rumelhart, Hinton, & Williams, 1988) was coupled to propose a general trainable model not only limited to linearly separable problems. The effective training using backpropagation made the usage of neural networks applicable to train sequential data using recurrent neural networks (Rumelhart et al., 1986) as well as digit recognition in computer vision (LeCun, Bottou, Bengio, & Haffner, 1998).

The emergence of graphical processing units (GPU) in 1999 and their ability to speed up the computation of huge matrices led to the first large neural network called AlexNet (Krizhevsky et al., 2012) in the field. After 2012, the area witnessed plenty of successful models under the rise of deep models, from beating human performance on the task of object recognition (He et al., 2016) using a large CNN model, success in medical image segmentation (Ronneberger, Fischer, & Brox, 2015) to conquering the world champion in the game of Go using reinforcement learning (Silver et al., 2017).

3.2 Arising of GANs

In addition to the success of discriminative approaches, generative approaches in both contexts of text and image also received attention. After introducing GAN in 2014, many studies considered using GAN instead of other familiar generative models at the time. Unlike most previous generative models, no Markov chains and inference is needed for GANs, which makes them very popular for various applications.

In the original paper of GAN, Goodfellow et al. (2014) used the idea of adversarial training to create realistic images. Two years later, Radford, Metz, and Chintala (2016) addressed the instability problem of GANs by introducing DCGAN and not only improved the quality of the generated images but also showed that GANs can learn meaningful representations and interpolating them can generate similar images in the current manifold. To further increase the quality of created images, Karras, Laine, and Aila (2019) proposed StyleGAN, a new generator model different from vanilla GAN by mapping the input noise to a new intermediate latent space and then passing it along with the initial noise through a gate to each convolution layer of the generator. Aside from generating realistic natural images, Han et al. (2018) utilized DCGAN and Wasserstein GAN (WGAN) (Arjovsky, Chintala, & Bottou, 2017) for medical image generation.

GANs were not limited to image generation and have been investigated for various applications. Isola, Zhu, Zhou, and Efros (2017) introduced Pix2Pix to propose an image to image translation using conditional GAN (Mirza & Osindero, 2014). While Zhu et al. (2017) proposed CycleGAN, an image-to-image translation model to go from a source domain to a target domain and vice versa. In another work, Choi et al. (2018) proposed StarGAN, a multi-domain image-to-image translation framework, by reconstructing the original image from the fake image. Aside from image-to-image translation models, there was a great deal of work considering GAN for text-to-image translation. In one of the first works, S. Reed et al. (2016) utilized a DCGAN conditioned on an embedded text description to generate pixels from characters. S. E. Reed et al. (2016) in the same year, introduced GAWWN to generate images with pre-defined content in a location it is asked for given textual instruction. In another work, H. Zhang et al. (2017) introduced StackGAN, a double-stage GAN model, to generate photo-realistic images conditioned on text descriptions. Dash, Gamboa, Ahmed, Liwicki, and Afzal (2017) suggested that using class conditional information would generate more diverse images in text-to-image translation tasks.

Rather than translating from one domain to another domain using adversarial training, GANs have been used in many other interesting applications. Antipov, Baccouche, and Dugelay (2017) proposed Age-cGAN conditioned on a required age category with a latent vector optimization approach to reconstruct an input face image preserving the original person's identity. Z. Zhang, Song, and Qi (2017) used a conditional adversarial autoencoder (CAAE) to smoothly get the progression and regression of the given image. Ledig et al. (2017) used GAN with the generator equipped with residual blocks for single image super-resolution (SISR) task. In another interesting work based on the idea of adversarial training, Pathak, Krahenbuhl, Donahue, Darrell, and Efros (2016) proposed an autoencoder training based on adversarial loss grasped from GAN's idea for photo inpainting. Yeh et al. (2017) introduced a semantic photo inpainting model based on a trained GAN on realistic images and then finding the missing part of the image by iteratively updating z to find the closest mapping on the latent image manifold. In (Vondrick, Pirsiavash, & Torralba, 2016), a GAN model with two generative paths of foreground and background along with a mask of motion pathway was used for video prediction. In 2016, Wu, Zhang, Xue, Freeman, and Tenenbaum (2016) proposed 3D-GAN to generate 3D objects. Gadelha, Maji, and Wang (2017) introduced PrGAN, a 3D GAN with a projection module, to generate 2D images from 3D shapes.

In addition to possible applications of GANs, they gained attention in medical imaging tasks. Nie et al. (2017) suggested an adversarial model to generate MRI images from CT images. Wolterink, Leiner, Viergever, and Išgum (2017) proposed an adversarial model by training two separate generators with two different losses and combining them to reduce the noise in low-dose CT images. Image super-resolution using GANs has also been investigated in medical images (Mahapatra et al., 2019; Sood, Topiwala, Choutagunta, Sood, & Rusu, 2018). The first model for anomaly detection using GAN was proposed by (Schlegl et al., 2017) with a similar idea as (Yeh et al., 2017).

3.3 Distribution Shift

Detecting and addressing different types of distribution shifts has a long history and spans from anomaly detection to domain adaptation. Domain adaptation approaches directly adapt to specific shifts, focusing more on covariate shift. On the other hand, anomaly detection makes a strong connection to distribution shifts, indicating a deviation from the expected norm. Unlike domain adaptation, where a model is tailored to recognize the type of shifts and further tackle them, anomaly detection only indicates a deviation in the observed data without specifying the shift.

3.3.1 Anomaly Detection

Detecting distribution shift can be simplified to anomaly detection, the task of identifying whether a single sample comes from the same distribution as the seen data.

Anomaly detection (AD) or, in general, out-of-distribution (OOD) detection has a long history in machine learning. It has been widely investigated in many applications, from network intrusion to medical diagnostics. AD approaches can be grouped according to three major paradigms.

Distributional-based approaches: The methods in this category try to build a probabilistic model on the distribution of normal data. They rely on the idea that the anomalous samples would act differently than the normal data. They expect that the anomalous samples receive a lower likelihood under the probabilistic model than the normal samples. The difference in these models is in the choice of their probabilistic model and their feature representation approach. Gaussian mixture models (Parzen, 1962), which only work if the data can be modeled with the probabilistic assumptions of the model, and kernel density estimation (KDE) (Latecki, Lazarevic, & Pokrajac, 2007) methods are among traditional methods. Some recent approaches use deep learning to represent the features, for instance, Zhou and Paffenroth (2017) introduced an autoencoder model based on robust principal component analysis (RPCA) to detect anomalies, and B. Yang, Fu, Sidiropoulos, and Hong (2017) proposed a combination of a deep neural network as dimensionality reduction and an SVM to cluster the data. To alleviate the limitation that the probabilistic assumption imposes, recent studies suggested learning a probabilistic model on the features extracted by the deep models such as DAGMM, which applies a Gaussian mixture model on input representations obtained from a deep autoencoder (Zong et al., 2018).

Classification-based approaches: One-Class SVM (OC-SVM) (Schölkopf et al., 1999) and support vector data description (SVDD) (Tax & Duin, 2004) are among the first works in this category. They used the idea of separating the normal data from the anomalous data based on their feature spaces. While the former is a kernel-based method (typically RBF kernels) trying to learn different sets forming the input space, the latter indicates anomalies by creating a spherically shaped boundary around the training data, then those residing outside these sets/boundaries are identified as anomalies. In the long history of the studies of this paradigm, different approaches, from kernel methods to deep learning approaches such as Deep-SVDD (Ruff et al., 2018) have been used. However, these approaches may suffer from the insufficient and biased representations the feature learning methods can provide. One remedy for this issue is using self-supervised learning methods. Various surrogate tasks such as image colorization (R. Zhang, Isola, & Efros, 2016), video frame prediction (Mathieu, Couprie, & LeCun, 2016), and localization (C. Yang, Wu, Zhou, & Lin, 2021) are among those that provide high-quality feature representations for downstream tasks. In 2018, Golan and El-Yaniv (2018) proposed geometric transformation classification (GEOM) to predict different geometric image transformations as their surrogate task for anomaly detection. Following that, Bergman and Hoshen (2020) introduced GOAD, a unified method of one-class classification and transformation-based classification methods. Most recently, Sohn, Li, Yoon, Jin, and Pfister (2021) presented a two-stage framework with a self-supervised model to obtain high-level data representations as the first stage, followed by a one-class classifier, such as OC-SVM or KDE, on top of the representations of the first stage.

Reconstruction-based approaches: Another approach to targeting anomalies is reconstruction-based methods. Instead of relying on the lower likelihood of the distributional-based methods, these approaches rely on the idea that normal samples should receive smaller reconstruction loss rather than anomalous samples. Different loss and reconstruction basis functions vary in each of these approaches. Previous studies have shown the advantages of both traditional machine learning models and deep neural networks as the basis reconstruction function, e.g. K-means in (Jianliang, Haikun, & Ling, 2009) and variational autoencoder in (An & Cho, 2015). In the class of deep neural networks, generative models such as GANs (Schlegl et al., 2017) and autoencoder (Zhou &

Paffenroth, 2017) are used to learn the reconstruction basis functions. Following the presentation of AnoGAN (Schlegl et al., 2017) as the first anomaly detection model based on GAN, several other studies used similar ideas with modifications on their basis functions and losses (Deecke, Vandermeulen, Ruff, Mandt, & Kloft, 2018; Rafiee & Fevens, 2020; Zenati, Foo, Lecouat, Manek, & Chandrasekhar, 2018; Zenati, Romain, Foo, Lecouat, & Chandrasekhar, 2018) to increase the performance of anomaly detection models based on GANs. Despite the increasing popularity of generative models, especially GANs, as the basis functions for anomaly detection, they are limited mainly by the two major issues of GANs, mode-collapse and catastrophic forgetting, putting their performance far behind classification-based approaches.

3.3.2 Domain Adaptation

Neglecting the existence of distribution shift under the i.i.d. assumption impacts many machine learning models, lowering their performance. Knowledge transfer from one domain to another under the possibility of changes in the domain emerged to address this issue. From the family of transfer learning approaches, domain adaptation is widely used to investigate and tackle different types of shifts.

Given various criteria, transfer learning approaches can be divided into different categories. Considering the importance of source data in the transfer mechanism as the main criterion, one can divide them into two categories, i.e. data-driven and model-driven approaches. Instance weighting models that assign weights to the source domain instances to reduce the marginal distribution differences between source and target domains introduce classical data-driven approaches (Bickel, Brückner, & Scheffer, 2007; Zadrozny, 2004). Deep adaptive network (DAN) (Long, Cao, Wang, & Jordan, 2015) was the first to use a convolutional neural network (CNN) with multi-kernel MMD (maximum mean discrepancy) for domain adaptation. Later, Ganin and Lempitsky (2015) proposed a domain adversarial neural network (DANN), a new domain adaptation model based on adversarial training. Recently, Y. Zhang, Liu, Long, and Jordan (2019) suggests margin disparity discrepancy(MDD) to measure the distribution discrepancy between domains. Despite the differences, all of these models have one thing in common, i.e. they all assume to have direct access to the source data during the knowledge transfer. To mitigate transfer learning models' privacy and storage concerns, source-free domain adaptation (SFDA) approaches (also known as model-driven or hypothesis transfer learning) are proposed. SFDA is a transfer learning strategy where a model trained on the source domain incorporates the learning procedure of the target domain. It was first introduced by Kuzborskij and Orabona (2013) where the access to the source domain was only limited to a set of hypotheses induced from it, unlike domain adaptation, where both source and target domains are used to adapt the source hypothesis to the target domain.

Source-free Domain Adaptation: In vision applications, Source-free Domain Adaptation (SFDA) has emerged as a learning strategy to adapt the knowledge learned from one dataset (source domain) to a query dataset (target domain) and has shown superior performance to more conventional domain adaptation methods. To this day, SFDA has mostly been experimented with covariate shift and assumes similar class distributions between source and target domains, i.e. no label distribution shift. Kuzborskij and Orabona (2013) introduced the idea of SFDA, also known as hypothesis transfer learning (HTL), and provided theoretical supports for its application under regularized least squares. Following that, Kuzborskij and Orabona (2017); X. Wang and Schneider (2015) studied the possibility of applying SFDA to other ML algorithms. Beyond theoretical analyses of SFDA, few studies focused on more general frameworks. Fernandes and Cardoso (2019) suggested a regularization approach to minimize the structural distance between source and target models. Although these works alleviate privacy and storage concerns related to typical domain adaptation approaches, they all assume access to a set of labeled data in the target domain. Liang, Hu, and Feng (2020) present SHOT, the first SFDA model with access to unlabeled target data. SHOT trained using mutual information maximization between the target hypothesis and the target data. In order to account for unlabeled data in the target domain, they suggested using a pseudo-labeling strategy. Similarly, Lao, Jiang, and Havaei (2021) proposed a new SFDA method with multiple hypotheses in source and target with a similar assumption as SHOT.

In contrast to covariate shift that has been largely investigated under the umbrella of domain adaptation and SFDA, label distribution shift between domains which is a common phenomenon in many real-world applications such as medical imaging has only been scarcely studied in the literature. Among the limited attempts at tackling label shift, either estimating the ratio between the two marginals, i.e. $p_S(y)/p_T(y)$, or the label proportions become the dominant strategy (Azizzadenesheli, Liu, Yang, & Anandkumar, 2019; Li, Murias, Major, Dawson, & Carlson, 2019; Lipton, Wang, & Smola, 2018; Redko, Courty, Flamary, & Tuia, 2019). K. Zhang, Schölkopf, Muandet, and Wang (2013) was one of the first to address target shift via a kernel mean matching method. Yet, their approach is not computationally applicable to larger datasets. Lipton et al. (2018) introduced Black Box Shift Estimation (BBSE) to estimate the importance weights using the confusion matrix. Recent practices estimate the ratio or proportions using optimal transport (OT), exploring the space of transport functions from source to target domain to find one with a minimum cost. MARS (Rakotomamonjy et al., 2022) relies on optimal transport to learn domain-invariant representations with sample re-weighting. In 2022, Kirchmeyer, Rakotomamonjy, Bezenac, and gallinari (2022) proposed OSTAR, a reweighing model which maps pretrained representations using optimal transport. Despite the effectiveness of these works, they have either one or both of the following assumptions; i) access to the source data during knowledge transfer, ii) their access to a set of labeled target set. Therefore, they are not feasible for source-free domain adaptation models.

Transfer Learning Application in Medicine: Transfer learning has been widely used in the medical field. Various studies suggested using pre-trained models on large natural image datasets such as ImageNet for medical imaging tasks; chest X-ray classification (Abbas, Abdelsamea, & Gaber, 2020) and brain tumor classification (Swati et al., 2019). The benefits of using ImageNet pre-trained models on medical imaging tasks have been inconclusive (Raghu, Zhang, Kleinberg, & Bengio, 2019). It has been shown to help in very small data regimes (Esteva et al., 2017; Raghu et al., 2019) but could hurt in case of datasets with ample data examples (Raghu et al., 2019). Moreover, the features learned by pre-training on ImageNet are not rich enough for 3D medical images (Litjens et al., 2017). Recent studies applied domain adaptation approaches to mitigate the distribution shift while benefiting the existing medical data. Among the works on medical imaging, several studies considered working on MR images as the source domain and CT images as the target domain using adversarial training to generate synthetic CT images (Abbas et al., 2020; Ouyang, Kamnitsas, Biffi, Duan, & Rueckert, 2019), cardiac structure segmentation (Dou, Ouyang, Chen, Chen, & Heng, 2018), and image registration (Mahapatra & Ge, 2020). One of the limitations in domain adaptation models, either supervised or unsupervised, is their necessity to access the source and target domain simultaneously, which is not always applicable in medical imaging tasks. To this end, recent studies investigate the application of hypothesis transfer learning in the medical imaging domain. Yu et al. (2020) proposed to utilize supervised SFDA to transfer knowledge from the source hypothesis to the target domain under knowledge distillation settings. Whereas in (X. Guo et al., 2020), SFDA has been applied to a federated learning setting to preserve patient privacy in diabetes prediction. X. Liu, Xing, Yang, El Fakhri, and Woo (2021) proposed one of the very first unsupervised SFDA models in the medical domain based on adaptive batch-wise normalization.

Ensemble Models: The idea of ensemble models goes back to (Dasarathy & Sheela, 1979) that divided the feature space with few classifiers. However, Schapire (1990) was the first who introduce ensemble models into machine learning through boosting. Ensemble in terms of majority voting among the prediction of several models as decision trees introduced in random forest (T. K. Ho, 1995). Ensemble learning made its way into many practical applications and machine learning competitions (Ayerdi, Savio, & Graña, 2013; Louzada & Ara, 2012; Mu, Lu, Watta, & Hassoun, 2009; Sill, Takács, Mackey, & Lin, 2009).

Beyond the overall improvement, ensembles have also been influential in neural network calibration. Lakshminarayanan et al. (2017) conducted a series of experiments on ensemble models in neural networks and showed that deep ensembles are the best-calibrated uncertainty estimators. An ensemble model, by definition, is the ability to correct the mistakes of its members. However, the most important factor in the success of an ensemble model lies in the diversity of its members. After all, if all members provided the same output, correcting a possible mistake would not be possible. Ensemble diversity has been widely investigated in the literature (Brown, 2004; Y. Liu & Yao, 1999). Improving diversity in neural network ensembles has become a focus in recent work. Stickland and Murray (2020) suggest augmenting each member of an ensemble with a different set of augmented input to increase the diversity among members. While a few recent studies propose deep ensemble models based on different neural network architectures to ensure diversity (Antorán, Allingham, & Hernández-Lobato, 2020; Zaidi et al., 2020). Recently Pagliardini, Jaggi, Fleuret, and Karimireddy (2022) suggest that encouraging diversity between the ensemble predictions helps to generalize in the OOD setting by increasing disagreements and uncertainties over out-of-distribution samples. Y. Lee, Yao, and Finn (2022) introduced an ensemble of multiple hypotheses with shared feature extractors and separate classifier heads to generalize in the presence of spurious features. They proposed to increase diversity among the classifiers through mutual information minimization over the hypotheses' predictions on unlabeled target data.

Chapter 4

Unsupervised Anomaly Detection with a GAN Augmented Autoencoder

In this chapter, we focus on detecting distribution shifts through anomaly detection. Specifically, we study the effect of generative adversarial networks on identifying anomalies. In order to facilitate detection with lower inference time, we investigate combining an autoencoder with a generative adversarial network.

4.1 Introduction

Anomaly detection (AD), or sometimes novelty detection, outlier detection, or in a broad description out-of-distribution detection, is an interesting and well-known research topic that is widely studied in many fields such as network intrusion (Leung & Leckie, 2005), fraud detection (Fawcett & Provost, 1997), and computer vision (Mahadevan, Li, Bhalodia, & Vasconcelos, 2010). The problem focuses on identifying samples that deviate from other observations on data, indicating variability in measurement, experimental errors, or a novelty. In other words, finding those samples that do not fit the training data distribution is known as anomaly detection. This can be helpful to identify unknown anomalies in the medical domain where finding an appropriate annotated dataset is always a concern. This approach is also applicable in cases where the knowledge regarding the type of anomalies is limited.

A generative adversarial network (GAN) (Goodfellow et al., 2014) has two components, a generator and a discriminator, with a multi-objective optimization which forms a zero-sum game between these two components leading to rich representations of the training data where these representations can be further utilized for downstream tasks. Generating realistic images of natural images (Karras et al., 2021, 2019, 2020; Radford et al., 2016) and medical images (Han et al., 2018), image-to-image translation (Isola et al., 2017; Zhu et al., 2017), and text-to-image translation (Dash et al., 2017; Lao et al., 2019; S. E. Reed et al., 2016; H. Zhang et al., 2017) are some of the recent practices that achieved state-of-the-art performance using the idea of GAN. Aside from the fact that GANs can model the training distribution, using them to identify anomalies requires the corresponding latent representation of a given test image which is not obtained easily. Previous studies suggested either optimizing the input noise to the GANs (Schlegl et al., 2017) or using another module trained alongside the GAN (Zenati, Foo, et al., 2018; Zenati, Romain, et al., 2018) to obtain the desired representation.

Following the importance of detecting anomalies in both natural and medical images, we present a simple and effective model based on GANs. In this model, a GAN and an autoencoder train simultaneously to learn the desired representations of the normal samples, which further will be used to indicate anomalies. In this work, anomalies are detected based on a new scoring function–a modification on previous anomaly score by considering multiple representations of a single image obtained from a GAN and an autoencoder. The experimental results on various domains; natural images (MNIST, CIFAR10, and SVHN), and medical imaging (Acute Lymphoblastic Leukemia (ALL) Labati, Piuri, and Scotti (2011)) datasets demonstrate that our suggested generative model is capable of identifying anomalous (out-of-distribution) samples in different settings. Our model not only improved all the existing models in all the experiments but also showed that even if it trained on a very small dataset, the representations are rich enough to target anomalies.

4.2 Related Work

There are numerous different approaches in the literature to identify anomalies in various domains. In the context of images, these studies can be divided into three sub-categories. 1) The first category of research considers classical machine learning (ML) approaches such as one-class support vector machines (SVMs) (Tax & Duin, 2004) and clustering (Xiong, Póczos, & Schneider, 2011) to detect anomalies. 2) The second type of work, also known as hybrid models, combine the classical ML and deep learning models; e.g., a one-class SVM on top of deep belief networks (DBNs) (Erfani, Rajasegarar, Karunasekera, & Leckie, 2016) or an autoencoder with a k-means clustering on top (Aytekin, Ni, Cricri, & Aksu, 2018). 3) The last category includes recent develops in deep learning and designed purely based on the representations they provide. Variational autoencoders (An & Cho, 2015) and autoencoders (Zhou & Paffenroth, 2017) showcase the power of deep models for detecting anomalies.

In the last category, there is a series of work that has been leveraging GANs to obtain the desired representations for the purpose of detecting anomalies. However, finding meaningful representations of the distribution of the normal images is a challenging task. In one of the very first works, Schlegl et al. (2017) proposed AnoGAN, a vanilla GAN accompanied with an optimization process on latent representation during inference procedure, to detect anomalies in the medical domain. A year later, Zenati, Foo, et al. (2018); Zenati, Romain, et al. (2018) proposed two different models based on BiGAN (Donahue, Krähenbühl, & Darrell, 2017), the recently proposed feature learning model, for the task of anomaly detection with a significant improvement on the inference time.

Following the recent successes using GANs and their variations on AD tasks, we introduce an unsupervised model based on GAN. Our model contains two generative models, a GAN and an autoencoder, to obtain the desired representation of a given image with two purposes, improving the performance of existing unsupervised AD models, and decreasing the detection time.

4.3 Anomaly Detection

The idea of using a GAN to find anomalies can be divided into two steps; learning the corresponding latent representation of a given image and a distance metric on how far the generated output is from the given image.

Previous studies each took advantage of GANs differently with a tailored distance metric for their proposed model to identify anomalies. In the next section, the similarities and differences of each of these two steps in the previous GAN-based models will be briefly described. Later on, the details of our AD model will be explained.

4.3.1 GANs for Anomaly Detection

Following the success of GANs and their application in various domains, Schlegl et al. (2017) introduced the first anomaly detection model based on GANs called AnoGAN. A GAN was trained on training data to learn the distribution of normal medical images which later can be used to target anomalous samples. To do so, an optimization process on the random noise to find the closest generated image to the input test image was proposed. They defined a distance metric to measure how well a given test sample is generated as a way to discriminate anomalies. Albeit the model showed that a vanilla GAN could discriminate normal images from anomalous, it imposed considerable computation on the model leading to a very slow inference process.

A year later, Zenati, Foo, et al. (2018) presented an unsupervised model based on a bidirectional generative adversarial network (BiGAN) model (Donahue et al., 2017; Dumoulin et al., 2017) with a similar scoring function as Schlegl et al. (2017) to accelerate the inference procedure[1].

Following the previous work, Zenati et al. proposed Adversarially Learned Anomaly Detection (ALAD) (Zenati, Romain, et al., 2018), a modification of their previous work, to detect anomalies. Their model contains three discriminators each receiving an input pair–one for handling the latent representations (D_{zz}), one for the input image x (D_{xx}), and D_{xz} which is similar to the discriminator used in BiGAN. For the inference, the L_1 reconstruction error in the feature space was used as the anomaly score:

$$A(x) = \|f_{xx}(x, x) - f_{xx}(x, G(E(x)))\|_1 \tag{7}$$

where f_{xx} is the activation of the layer before the logits in the D_{xx} network, $E(x)$ is the representation obtained from the encoder E for the given image x, and $G(E(x))$ is the output of the generator G given $E(x)$.

[1] For simplicity, we refer it as Efficient-GAN in the experiments and results section.

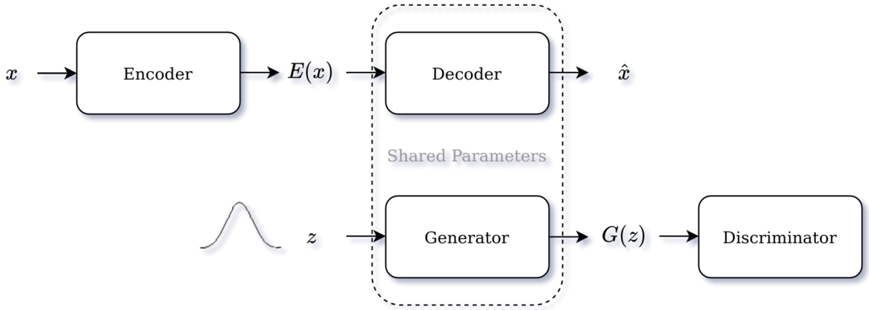


Figure 4.1: The GAN and autoencoder used in our model; encoder and discriminator have similar architecture except in their last layers, and the generator and the decoder share their weights.

4.3.2 Our Anomaly Detection Model

Similar to the previous AD models based on GANs, we suggest using adversarial training to identify anomalies. We present a generative model, a combination of a GAN and an autoencoder (see Fig. 4.1). In this setting, we use parameter sharing (also known as weight sharing) between the GAN's generator and autoencoder's decoder to keep their distribution as close as possible. This will benefit the inference process by helping the encoder to generate representations within the distribution of the GAN. Our AD model trains on $D_{ind} = \{x_1, x_2, ..., x_k \sim P_{ind}\}$ where P_{ind} defines normal (in-distribution) training samples. Therefore, the generated outputs of the GAN and the encoded representation of the encoder will be close to P_{ind}. During the inference, the model tests on $D_{mix} = \{x_1, x_2, ..., x_k \sim P_{ind} \; or \; P_{ood}\}$ where P_{ood} defines anomalous (out-of-distribution) samples. Hence, the expected outputs of the GAN and the encoded representation of the autoencoder for an anomalous sample will be far from the actual test image and in other words close to P_{ind}. As a result, the dissimilarity between a given test sample and its corresponding generated output can be defined as our distance metric to target anomalous samples.

We train the GAN with relativistic standard GAN (RSGAN) (Jolicoeur-Martineau, 2019) loss. Unlike the standard GAN (SGAN) objective function which measures the probability that the input data is real, Relativistic GAN measures the probability that the real data is more realistic than the generated data (or vice versa).

$$\begin{aligned} L_D^{RSGAN} &= -\mathbb{E}_{(x_r, x_f)\sim(\mathbb{P},\mathbb{Q})}[\log(\mathrm{sigmoid}(C(x_r) - C(x_f)))] \\ L_G^{RSGAN} &= -\mathbb{E}_{(x_r, x_f)\sim(\mathbb{P},\mathbb{Q})}[\log(\mathrm{sigmoid}(C(x_f) - C(x_r)))] \end{aligned} \tag{8}$$

where G and D are the generator and discriminator of the GAN, $\mathbb{P}$ is the distribution of the real data, $\mathbb{Q}$ is the distribution of the fake data, x_r and x_f are real and fake data, and C is the critic.

The autoencoder AE was trained using the mean squared error (MSE) reconstruction loss function, $L_{AE} = \|x - G(E(x))\|^2$, where $E(x)$ is the encoded representation of an input image x produced by encoder E.

The anomaly score presented in this work modifies the previous scoring function presented in (Schlegl et al., 2017).

$$A(x) = \lambda L_D(x) + (1 - \lambda) L_R(x) \tag{9}$$

As it is shown in Eq. 9, in (Schlegl et al., 2017), the anomaly score of image x, $A(x)$, includes two terms–discrimination loss, $L_D(x)$, and residual loss, $L_R(x)$. These two terms compute the difference between the actual test image and its corresponding generated output from two different perspectives. $L_D(x)$ relies on the discriminator's intermediate representations ($f_D(\cdot)$) for them (Eq. 10), while the $L_R(x)$ computes their visual dissimilarity (Eq. 11).

$$L_D(x) = \sum |f_D(x) - f_D(G(E(x)))| \tag{10}$$

$$L_R(x) = \sum |x - G(E(x))| \tag{11}$$

As stated earlier, we consider multiple representations of a single image to identify anomalies. Therefore, rather than discrimination loss and residual loss, we suggest using the encoded representation of the encoder as the latent loss, L_L (shown in Eq. 12). For a given image x, $L_L(x)$ computes how far the encoded representation of x, $E(x)$, is from the encoded representation of its generated output given $E(x)$.

$$L_L(x) = \sum |E(x) - E(G(E(x)))| \tag{12}$$

By adding the latent loss to Eq. 9, we present a new anomaly score function, given in Eq. 13. The effect of latent loss in our scoring function is controlled by the hyperparameter β.

$$A(x) = \lambda L_D(x) + (1 - \lambda) L_R(x) + \beta L_L(x) \tag{13}$$

Given the anomaly score presented here, once the model learns the true distribution of normal (in-distribution) samples, P_{ind}, it will identify anomalous (out-of-distribution) samples, P_{ood}, by a higher anomaly score assigned to them as opposed to the score for the normal samples.

4.4 Datasets

To evaluate the performance of our model on AD tasks, in comparison with recent GAN-based models, we considered two types of datasets–natural images and medical images. MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky, 2009), and SVHN (Netzer et al., 2011) as three benchmarks for natural images were chosen. For the medical dataset, we considered the Acute Lymphoblastic Leukemia (ALL) dataset (Labati et al., 2011) with only 260 images to evaluate our model's capability to perform under a limited data regime, which is quite common in the medical domain.

Unlike the medical dataset, which provides normal and anomalous classes, each of the natural datasets has 10 classes. Therefore each of those classes/labels separately can be defined as either normal or anomalous for our AD task. To this end, two new strategies to form the new datasets from the natural image datasets have been introduced here: 1) we define *1 versus 9* where one out of 10 classes is chosen to be anomalous while the rest form normal class, and 2) *9 versus 1* where nine classes form the anomalous class and the remaining one form the normal class. These two strategies create 20 different datasets for each of the natural datasets, with a total of 60 datasets.

In the experiments on natural images, only normal images are considered for the training, while anomalous images and test data are used for the inference. In these experiments, a small proportion of samples is used as the validation sets. In order to evaluate the model on another domain with a fewer number of samples, the Acute Lymphoblastic Leukemia (ALL) dataset, with 260 samples and an equal number of normal and anomalous samples for each class, is considered. From D_{ind}, 100 samples are used for training, 20 samples for validation, and the remaining 140 samples from

D_{mix}, are considered for evaluation.

4.5 Experiments and Results

The proposed model's performance was evaluated on natural (MNIST, CIFAR10, and SVHN) and medical (ALL) images. To be able to determine our model's benefits as well as its weaknesses, a comparison has been made on similar GAN-based AD models, Efficient-GAN (Zenati, Foo, et al., 2018), ALAD (Zenati, Romain, et al., 2018), and AnoGAN (Schlegl et al., 2017). Except for the AnoGAN, which suffers from a very long inference procedure (see Sec. 4.5.2), all the other models were evaluated on all four datasets.

The detailed information of the choices of hyperparameters for our model on each of the experiments is indicated in Table 5.1. For the medical domain, similar hyperparameters are used for all the GAN-based models compared in this study, while for the natural images, we used similar hyperparameters as presented in (Zenati, Foo, et al., 2018; Zenati, Romain, et al., 2018). In the case of the SVHN dataset, we compared our model with AnoGAN and ALAD following similar hyperparameters as (Zenati, Romain, et al., 2018).

For the experiment on the medical dataset, we trained each model for 1000 epochs on D_{ind} with a learning rate of 1e$-$4, batch size of 16, latent size of 200, and dropout ratio of 0.2 for the encoder and discriminator. The models trained on natural image datasets for at most 85 epochs, batch size of 64, and learning rate of 1e$-$4. The latent sizes of 100 for MNIST and CIFAR10,

Table 4.1: The architecture and hyperparameters of our model for the experiments on the MNIST, CIFAR10, SVHN and ALL datasets; the generator of the GAN and decoder of autoencoder use weight sharing. We used $i = 0, 0, 0, 1$, $j = 3, 4, 4, 8$, $k = 2, 0, 0, 0$, $l = 3, 3, 3, 6$, $m = 2, 1, 1, 1$, $p = 3, 4, 4, 6$, and $q = 1, 0, 0, 1$ for MNIST, CIFAR10, SVHN and ALL dataset respectively.

	Our model architecture		
Module	#Layers	Activation fn	Dropout
$G(z)$	**i** × $Conv2d$, **j** × $Trans.Conv2d$, **k** × $Linear$	$ReLU$	×
$D(x)$	**l** × $Conv2d$, **m** × $Linear$	$LeakyReLU$	0.2
$E(x)$	**p** × $Conv2d$, **q** × $Linear$	$LeakyReLU$	0.2
Learning rate	Lr_{GAN}: 1×10^{-4}, Lr_{AE}: 1×10^{-4}		
Optimizer	AdamW($\beta_1 = 0.5$, $\beta_2 = 0.999$)		
Batch Size	64 (except ALL with 16)		

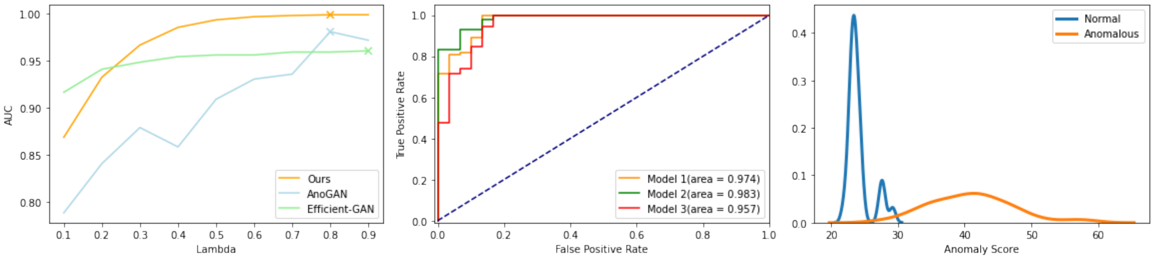


Figure 4.2: In the left, the performance of Efficient-GAN, AnoGAN, and Ours for different contributions of discrimination and residual losses under coefficient λ are depicted. In the middle and right, the performance of our model on three runs with their ROC curves along with the anomaly score distribution of our best models out of three runs are shown.

and 200 for SVHN were used, respectively. In all the experiments, models are optimized using the AdamW (Loshchilov & Hutter, 2019) optimizer. During the inference, different values of β were used for each dataset. These values were determined experimentally and defined the contribution of the latent loss in the new anomaly score. Specifically, $\beta = 1$ for CIFAR10, SVHN, and ALL datasets and $\beta = 0.5$ for MNIST dataset were used. $\lambda = 0.8$ was chosen experimentally for all the experiments.

4.5.1 Experimental Setup

The impact of λ

One of the key factors in the performance of the recent GAN-based models is the effectiveness of their scoring function. In (Schlegl et al., 2017; Zenati, Foo, et al., 2018) as well as our model, different contributions of the learned features of the critic (discrimination loss) and the visual dissimilarity of the generated samples and actual test samples (residual loss) in the final scoring can have a huge impact on the performance of each of these models. In a small experiment on the ALL dataset, the effect of different values of λ in the range of $[0, 1]$ on the performance of Efficient-GAN, AnoGAN, and our model was investigated. These models were compared based on their area under the ROC (receiver operating characteristic) curve (AUC). For the experiment on our modified scoring function, we used a fixed value of 1 for β. It can be observed from Fig. 4.2 that all these models perform better with larger λ, indicating a higher contribution of the discrimination loss.

Since the residual loss is more sensitive to the artifacts in the generated output, comparing

Table 4.2: The AUC (%) comparison on the ALL dataset for AnoGAN, Efficient-GAN, and our model with 0.8, 0.9, and 0.8 for coefficient λ for each method, respectively. In this and the following tables, the results obtained from our implementation are represented by the † sign. (± std. dev.)

Model	Sensitivity	Specificity	f1-measure	Accuracy	AUC
AnoGAN† (Schlegl et al., 2017)	73.08 ± 0.254	74.44 ± 0.164	79.19 ± 0.203	73.34 ± 0.236	75.71 ± 0.241
Efficient-GAN† Zenati, Foo, et al. (2018)	71.54 ± 0.229	**98.89** ± 0.016	81.07 ± 0.165	76.67 ± 0.183	87.23 ± 0.137
ALAD† (Zenati, Romain, et al., 2018)	94.61 ± 0.016	75.0 ± 0.057	88.52 ± 0.016	86.09 ± 0.022	79.88 ± 0.048
Ours	**98.72** ± 0.004	84.44 ± 0.016	**97.73** ± 0.001	**96.04** ± 0.003	**97.31** ± 0.009

the internal representation of a given image might ignore those visual differences and focus on more abstract features. More detail on the effectiveness and challenges of using these two losses is explained in the analysis section (Sec 4.5.2).

Stabilizing Training

One of the biggest challenges in training GANs is their instability. Slight changes in model's hyperparameters, running on different machines, and even random initialization can affect their performance more than any other deep models (Lucic et al., 2018). Therefore to reduce the instability of our model during training, spectral normalization (Miyato, Kataoka, Koyama, & Yoshida, 2018) was used for the critic. To compare the model's performance independent of its random initialization, the model was trained three times with different random initializations. All of the results reported in this study were computed as an average on the three runs from these different random initializations.

4.5.2 Experimental Results

Medical Imaging Dataset

The detailed performance of all four GAN-based models on our medical imaging dataset is summarized in Table 4.2. As illustrated in the table, our method showed a high capability to detect anomalies from various performance metrics. Ours outperformed the existing approaches on AUC with a large margin (increased by 10%). In terms of specificity, the best performance is acquired by Efficient-GAN with ours as the second best.

The observation on the range of standard deviation from multiple runs showed that AnoGAN had the least stability. In comparison, the highest stability is achieved by ours, which can be inferred from both ROC curves of ours on three runs (Fig. 4.2, middle plot) and the results from Table 4.2. We also showed that our model could effectively discriminate normal and anomalous samples even on a very small dataset (Fig. 4.2, third plot from the left).

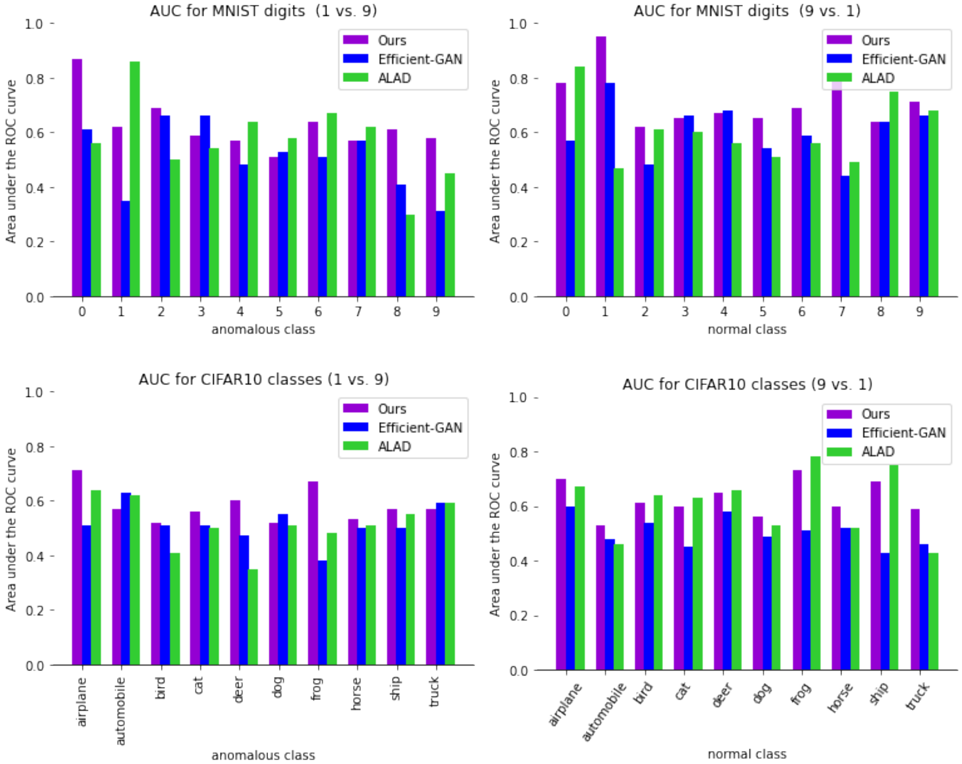


Figure 4.3: Individual performance of each label on MNIST and CIFAR10.

Natural Images

For our experiments on natural images, we considered the aforementioned *9 versus 1* and *1 versus 9* strategies and compared the performance of our model with Efficient-GAN (Zenati, Foo, et al., 2018) and ALAD (Zenati, Romain, et al., 2018). Table 4.3 summarizes the AUC of each model within each of these strategies, which are averaged over three runs on all the classes of MNIST and CIFAR10. As the results reveal, our model outperformed the other two GAN-based models on all the experiments by a large margin. The detailed results of all three compared models on each of the classes of MNIST and CIFAR10 are also depicted in Fig. 4.3.

Table 4.3: The AUC (%) comparison on MNIST and CIFAR10 datasets with *one-vs-all* and *all-vs-one* strategies. Results from the original papers are indicated by the * symbol. (± std. dev.)

	all-vs-one		*one-vs-all*	
Model	MNIST	CIFAR10	MNIST	CIFAR10
Efficient-GAN† (Zenati, Foo, et al., 2018)	50.9 ± 0.116	51.5 ± 0.064	60.4 ± 0.096	50.6 ± 0.053
ALAD† Zenati, Romain, et al. (2018)	57.2 ± 0.140	51.6 ± 0.086	60.7 ± 0.112	60.7 ± 0.120*
Ours	**62.5** ± 0.093	**58.2** ± 0.060	**71.6** ± 0.096	**62.6** ± 0.061

Table 4.4: The AUC (%) of AnoGAN, ALAD and our model on SVHN dataset with *all-vs-one* and *one-vs-all* strategies.

Model	*all-vs-one*	*one-vs-all*
AnoGAN (Schlegl et al., 2017)	46.6 ± 1.3	54.1 ± 0.019
ALAD (Zenati, Romain, et al., 2018)	51.6 ± 0.09	57.5*± 0.027
Ours	**56.8** ± 0.007	**58.1** ± 0.014

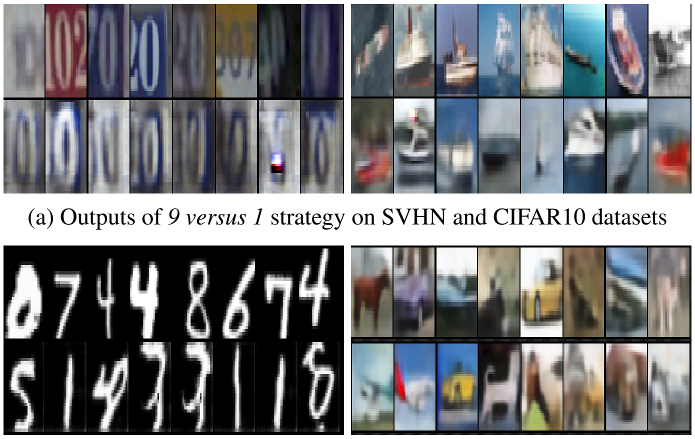
On the SVHN dataset, we evaluated the performance of our model on both *9 versus 1* and *1 versus 9* strategies. As shown in Table 4.4, our model outperformed its two other rivals on *1 versus 9* strategy with at least 5% improvement on AUC. The results however indicate a slight improvement (less than 1%) in the performance of our model on *9 versus 1* in comparison with ALAD.

Output Analysis

A thorough analysis of the generated outputs of our model on different datasets revealed that in the case of *9 versus 1* when only one of the labels form the normal class, the model is better able to capture the distribution of the normal data which is reasonable considering the model is learning an easier pattern. Even though this is the case for almost all the datasets (Fig. 4.4 (a) right; *9 versus 1* on CIFAR10), the model has difficulty when training on the SVHN dataset even when it should learn the distribution of just a single label representing the normal class. This is mostly due to the nature of this dataset where the classes are not completely separated, i.e., the samples of the class zero can contain other digits in their image (Fig. 4.4 (a) left) which makes it hard for the model to learn the true distribution of the digit zero. This phenomenon can affect the performance, especially during inference time where the visual dissimilarity of the generated image and the actual test image can have a direct impact on identifying anomalous samples. The model can easily fail and even if the model is able to generate the test digit, there can be often visual artifacts causing high residual loss.

We also observed that, in the cases where digits with similar patterns are considered as the normal class (with *9 versus 1* strategy), the model may fail to identify the anomalous image when the corresponding test image has a similar pattern. For instance, when considering digit 3 as the normal sample, the model can fail when the actual test image is digit 8, hence, receiving lower residual and discrimination losses and therefore will be recognized as a normal sample.

Considering *1 versus 9* strategy where 9 classes form the normal training data, mode collapse was the major issue in training the model for our anomaly detection purpose. As an example, in Fig. 4.4 (b), the model is more focused on learning the distribution of cars and planes in CIFAR10 dataset and digit seven and digit one while training on MNIST dataset and ignores the other classes. As the result, it may fail to learn the whole distribution while focusing on only a subset of the training distribution, therefore leading to high anomaly scores for the samples actually coming from the normal training distribution.



(a) Outputs of *9 versus 1* strategy on SVHN and CIFAR10 datasets

(b) Outputs of *1 versus 9* strategy on MNIST and CIFAR10 datasets

Figure 4.4: The generated outputs of our model on SVHN, CIFAR10, and MNIST datasets using *9 versus 1* and *1 versus 9* strategies. The top rows of each sub-figure (a) and (b) show the training images, and the second rows are the generated images by the GAN.

Inference Time Comparison

One of the major challenges in training a vanilla GAN for anomaly detection is its long inference process which negatively affects required time and computational resources for performance. Therefore, we modified the GAN by adding an autoencoder to help the model improve the existing

Table 4.5: Inference time comparison on the ALL dataset on images of shape (3, 220, 220) with a (200, 1) vector of noise randomized from a Gaussian distribution.

Models	# of parameters in each module			Inference time (ms)
	Encoder	Decoder/Generator	Critic	
AnoGAN† (Schlegl et al., 2017)	-	2,450,307	5,159,170	13110.47
Efficient-GAN† (Zenati, Foo, et al., 2018)	5,874,352	1,906,240	7,024,929	3.33
ALAD† Zenati, Romain, et al. (2018)	5,771,752	1,906,240	7,814,915	3.85
Ours	8,716,888	2,450,307	5,159,170	2.90

results while reducing the inference time. A comparison on all the GAN-based models studied in this work on the ALL dataset with 160 test images is shown in Table 4.5. As observed from the Table, Ours slightly improved the inference time compared to (Zenati, Foo, et al., 2018) and (Zenati, Romain, et al., 2018), while the improvement is more notable compared to (Schlegl et al., 2017). Python 3.7 with the PyTorch (Paszke et al., 2019) library on a GeForce GTX 1080 Ti GPU was used for these experiments. We considered 500 iterations for AnoGAN to optimize the random noise z for each given test image.

4.5.3 Ablation Study

Latent Loss Impact

The new anomaly score presented in this work is a modification of an existing scoring function (Schlegl et al., 2017; Zenati, Foo, et al., 2018) where we try to leverage the learned features of the autoencoder. Therefore to show the effectiveness of the new anomaly score, a comparison on the natural images using the new and original anomaly score was conducted. The results on Table 4.6 demonstrate the benefit of the added latent loss in the new anomaly score.

Table 4.6: The effect of latent loss in the new anomaly score. The comparison were done on natural images. In the experiments using latent loss, 0.5 and 1 were used as β for MNIST and CIFAR10 respectively. (± std. dev.)

	1 versus 9		*9 versus 1*	
Model	MNIST	CIFAR10	MNIST	CIFAR10
without latent loss ($\beta = 0$)	54.7 ± 0.099	50.2 ± 0.084	67.0 ± 0.114	56.9 ± 0.107
with latent loss ($\beta \neq 0$)	**62.5** ± 0.093	**58.2** ± 0.060	**71.6** ± 0.096	**62.6** ± 0.061

GAN Objective

To have a better understanding of the effectiveness of Relativistic GAN loss for our model, two different losses for GAN have been considered. Precisely, RSGAN and SGAN objective functions were compared on the natural datasets experimented on in this study. In all of the experiments, using RSGAN increased the performance of our model (see Table 4.7).

Table 4.7: The effect of using different GAN objective functions on the performance of our model. (± std. dev.)

	1 versus 9		*9 versus 1*	
GAN objective fn.	MNIST	CIFAR10	MNIST	CIFAR10
Standard GAN (SGAN)	55.7 ± 0.075	55.5 ± 0.082	69.3 ± 0.129	61.1 ± 0.088
Relativistic GAN (SRGAN)	**62.5** ± 0.093	**58.2** ± 0.060	**71.6** ± 0.096	**62.6** ± 0.061

4.6 Conclusion and Future Work

In this work, we suggested using a simple and effective generative model to identify anomalies. The model contains a GAN and an autoencoder, which train simultaneously on the normal training data. To detect anomalies during inference time, we introduced a new anomaly score function comprising multiple representations obtained from the autoencoder and the GAN. We further evaluated our model on MNIST, CIFAR10, SVHN, and a public Acute Lymphoblastic Leukemia (ALL) datasets. Our model proved its performance in all of the experiments with a large improvement over the existing GAN-based models with lower inference time. We also showed that our model could perform quite well even on small-sized datasets. Despite the effectiveness of our model on identifying anomalies, mitigating the challenges in training GANs and learning more complicated distribution seem to be necessary. To this end, in our future work, we tend to study the effect of using contrastive learning in training GANs to learn more discriminative representations of the images while investigating different scoring functions to fill this gap.

Chapter 5

AD-CGAN: Contrastive Generative Adversarial Network for Anomaly Detection

In this chapter, we continue our former research on using the generative adversarial network for anomaly detection. We investigate how integrating contrastive learning with the generative adversarial network benefits their ability for anomaly detection. This approach mitigates catastrophic forgetting and mode collapse of GANs simultaneously, hence, improving their performance in identifying anomalous data.

5.1 Introduction

Anomaly detection (AD), also known as out-of-distribution detection, has a long history in artificial intelligence. Anomaly detection refers to identifying those samples that do not come from the expected distribution. Supervised learning models address AD using classification approaches such as outlier exposure (Hendrycks, Mazeika, & Dietterich, 2019). On the other hand, unsupervised learning approaches, such as reconstruction-based methods (Schlegl et al., 2017; Zhou & Paffenroth, 2017), mitigate the problem of limited labeled data and unknown anomalies. In these approaches, the model learns the distribution of the normal training data and then a reconstruction loss targets anomalies. AnoGAN (Schlegl et al., 2017) proposes using generative adversarial networks (GANs) to find anomalies in the medical domain. AnoGAN suffers from its lengthy inference procedure to find the inverse mapping of an image in a low-dimensional representation. Several studies tried to overcome the limitations of AnoGAN (Rafiee & Fevens, 2020; Zenati, Foo, et al., 2018; Zenati, Romain, et al., 2018). However, intrinsic problems of GANs, such as mode collapse (Heusel et al., 2017), catastrophic forgetting (L. Chen et al., 2019; Kemker et al., 2018), unstable training, and difficulty in convergence (Lucic et al., 2018), limit the ability of these models to learn a suitable representation for the task of AD.

K. S. Lee, Tran, and Cheung (2021) showed that adding contrastive learning on the generator side in training GANs while maximizing mutual information on the discriminator side increases the quality of generated images by simultaneously mitigating mode collapse and catastrophic forgetting of the generator and the discriminator respectively. Contrastive learning (T. Chen, Kornblith, Norouzi, & Hinton, 2020; Hadsell, Chopra, & LeCun, 2006) is a self-supervised approach that learns representations of the data in such a way that similar samples stay close to each other while dissimilar samples remain at a distance. Considering K. S. Lee et al. (2021), we investigate the incorporation of a contrastive GAN for anomaly detection.

In this work, we propose a reconstruction-based Anomaly Detection approach using Contrastive Generative Adversarial Network (AD-CGAN). The proposed model contains three main sub-modules: a contrastive GAN, an autoencoder, and a second discriminator (different from the discriminator in GAN) on the latent representations. We train all modules simultaneously on the normal data to learn a discriminative representation for each image while keeping each image's local and global features as close as possible. The second discriminator trains on the hidden representations of two different reconstruction-based models, i.e., GAN and autoencoder, to provide more discriminative representations. We show that having a contrastive GAN while maximizing the mutual information between local and global features of an image provides more semantic and discriminative features for anomaly detection. Experimental results show that the representations obtained by the contrastive GAN in our anomaly detection model greatly increase the performance of reconstruction-based AD approaches. To the best of our knowledge, our work is the first to investigate using contrastive generative adversarial networks for anomaly detection.

5.2 Related Work

Anomaly detection or, in general, out-of-distribution detection approaches can be grouped according to the following paradigms.

Distributional-based approaches try to build a probabilistic model on the distribution of normal data. They expect that the anomalous samples receive a lower likelihood under the probabilistic model than the normal samples. Gaussian mixture models (Parzen, 1962) and kernel density estimation (KDE) (Latecki et al., 2007) from traditional models and RDA (Zhou & Paffenroth, 2017) and deep autoencoding Gaussian mixture model (DAGMM) (Zong et al., 2018) from deep models are among these approaches.

Classification-based approaches such as One-Class SVM (OC-SVM) (Schölkopf et al., 1999) and support vector data description (SVDD) (Tax & Duin, 2004) use the idea of separating the normal data from the anomalous data based on their feature spaces. These approaches suffer from the insufficient and biased representations the feature learning methods can provide. One remedy for this issue is to use self-supervised learning methods. GEOM (Golan & El-Yaniv, 2018) and GOAD (Bergman & Hoshen, 2020) are classification-based AD models that use surrogate tasks for anomaly detection.

Reconstruction-based approaches rely on the idea that normal samples should receive smaller reconstruction loss rather than anomalous samples. Various loss and reconstruction basis functions are used in each of these approaches. K-means is used as an early basis reconstruction function (Jianliang et al., 2009) while An and Cho (2015) proposed using deep neural networks as the basis functions. In the class of deep neural networks, generative models such as GANs (Schlegl et al., 2017) and autoencoder (Zhou & Paffenroth, 2017) are used to learn the reconstruction basis functions. Following the presentation of AnoGAN (Schlegl et al., 2017), several other studies used similar ideas with modifications on their basis functions and losses (Deecke et al., 2018; Rafiee & Fevens, 2020; Zenati, Foo, et al., 2018; Zenati, Romain, et al., 2018) to increase the performance of anomaly detection models based on GANs. One major issue in using generative models, especially GANs as the reconstruction basis function, is their difficulty to recover the entire data distribution (also known as mode-collapse in GANs), leading to lower performance in comparison with classification-based approaches. Our model falls in the category of reconstruction-based approaches, combining adversarial training with contrastive learning to mitigate the challenges of reconstruction-based approaches.

5.3 Background

Unsupervised anomaly detection models only have access to the normal training data. Reconstruction-based models are unsupervised approaches that rely on the reconstruction loss of samples, where a high reconstruction loss implies an anomalous sample. Our model uses a GAN and an autoencoder as its reconstruction methods.

Generative adversarial network (GAN) (Goodfellow et al., 2014) is a generative model that formulates the process of learning in a two-player minimax game between two learning components; i.e., generator and discriminator. One of the obstacles in using GANs for tasks such as anomaly detection is related to the catastrophic forgetting (neural network forgetting prior tasks while working on the current task) of the discriminator (T. Chen et al., 2019; Kemker et al., 2018) which can negatively affect the AD performance. Another barrier is known as mode collapse where the generator only learns a small subset of modes in the training data. Recently self-supervised learning has gained attention in generative models such as GANs (T. Chen et al., 2019; Tran, Tran, Nguyen, Yang, & Cheung, 2019). While these approaches try to mitigate the catastrophic forgetting, they do not diminish the mode collapse (Tran et al., 2019). On the other hand, maximizing mutual information on the discriminator side and contrastive learning on the generator side seems a way to overcome these two issues simultaneously (K. S. Lee et al., 2021). Therefore, we consider the idea of using a contrastive GAN in our AD model to detect anomalous samples with the purpose of increasing the performance of reconstruction-based models.

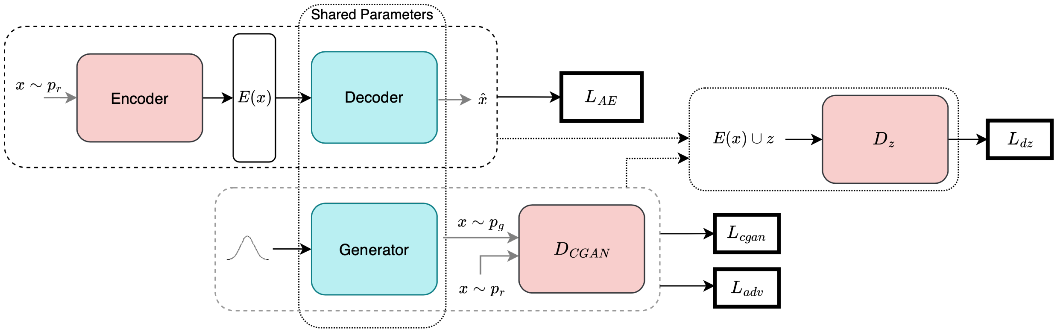


Figure 5.1: Different components and losses of AD-CGAN.

5.4 Proposed Approach

In this work, we propose AD-CGAN, a reconstruction model based on Contrastive Generative Adversarial Network to find anomalies in images (see Fig. 5.1). In this approach, a contrastive GAN module learns to generate normal samples. An autoencoder module, which shares its decoder with the GAN's generator, learns to reconstruct normal samples from their latent representations. We also use a discriminator module on top of the autoencoder and the input random noise to the GAN as a regularizer. Since the model only trained on normal samples, we expect that during inference, it cannot reconstruct samples from any distribution other than training distribution. As a result, the dissimilarity between a given test sample and its corresponding generated output can be defined as our distance metric to target anomalous samples. Therefore, we define a normality score based on the reconstruction loss of the input sample during inference to find anomalous samples. In the following sections, we discuss all the modules and the normality score in more detail.

5.4.1 Contrastive GAN

Formal Definition: The training set $\mathcal{D}_{train} = \{x_1, x_2, ..., x_k \sim P_{ind}\}$ contains samples drawn from P_{ind}, normal distribution. To evaluate our model, we use a test set $\mathcal{D}_{test} = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n \sim P_{ind} \cup P_{ood}\}$ including both normal and anomalous samples drawn from P_{ind} and the anomalous distribution (P_{ood}), respectively.

The contrastive GAN module which we refer to as $CGAN$ contains a generator G and a discriminator D_{cgan}. Training the $CGAN$ incorporates two losses: a contrastive loss L_{cgan} (See Eq. 14) and an adversarial loss L_{adv} (See Eq. 15).

In conventional contrastive learning, each image is contrasted with other samples, while in AD-CGAN, each image is contrasted with its own local feature maps to create positive/negative sets. Given an image $x \in X$, we consider the penultimate and ultimate representations of D_{cgan} as local ($C_\psi(x)$) and global ($E_\psi(x)$) features of x. We pass $E_\psi(x)$ through a dense layer ϕ_θ. Then, $\phi_\theta(E_\psi(x))$ and $C_\psi(x)$ go to the contrastive pairing phase to create positive/negative sets for the contrastive learning. For a given image x, the set of positive samples is the pair $(C_\psi^{(i)}(x), \phi_\theta(E_\psi(x)))$ for $i \in A = \{0, 1, ..., M^2 - 1\}$ of a $M \times M$ local feature map. Besides the local feature map of other images $x' \in X$ in the same mini-batch, we also consider the pairs $(C_\psi^{(j)}(x), \phi_\theta(E_\psi(x)))$ $j \in A,\ j \neq i$, from the same image x as the negative set. The contrastive loss of AD-CGAN, shown in Eq. 14, follows the loss presented in K. S. Lee et al. (2021) with a slight modification to fit the architectural design of our model:

$$\begin{aligned} L_{cgan}(X) &= -\mathbb{E}_{(x\in X)}\mathbb{E}_{(i\in A)}[\log\ p(C_\psi^{(i)}(x), E_\psi(x)|X)] \\ &= -\mathbb{E}_{(x\in X)}\mathbb{E}_{(i\in A)}\left[\log \frac{\exp(g_\theta(C_\psi^{(i)}(x), E_\psi(x)))}{\sum_{(x\prime,i)\in X\times A}\exp(g_\theta(C_\psi^{(i)}(x'), E_\psi(x)))}\right] \end{aligned} \tag{14}$$

Here the function $g_\theta(C_\psi^{(i)}(x), E_\psi(x)) = C_\psi^{(i)}(x)^T\phi_\theta(E_\psi(x))$ maps the local/global features with K dimensions to a scalar score. For the adversarial loss, we used relativistic loss (Jolicoeur-Martineau, 2019):

$$\begin{aligned} L_{adv_D} &= -\mathbb{E}_{(x_r\in X_r,\ x_f\in X_f)\sim(\mathbb{P},\mathbb{Q})}[\log(\sigma(C(x_r) - C(x_f)))] \\ L_{adv_G} &= -\mathbb{E}_{(x_r\in X_r,\ x_f\in X_f)\sim(\mathbb{P},\mathbb{Q})}[\log(\sigma(C(x_f) - C(x_r)))] \end{aligned} \tag{15}$$

where L_{adv_G} and L_{adv_D} are the losses of the generator and the discriminator of the CGAN, σ is the sigmoid function, X_r and X_f represent sets of real and fake images respectively, $\mathbb{P}$ is the distribution of the real data, $\mathbb{Q}$ is the distribution of the fake data, and C is the critic.

In order to stabilize training, we constrained the discriminator D_{cgan} and the generator to learn only from the contrastive loss of real image and fake image features, respectively, as suggested in (K. S. Lee et al., 2021). The final loss of the generator and the discriminator of our $CGAN$ is a combination of its adversarial and contrastive losses where α and β control the contribution of the contrastive loss:

$$\begin{aligned} L_G &= L_{adv_G} + \alpha L_{cgan}(X_f) \\ L_{D_{cgan}} &= L_{adv_D} + \beta L_{cgan}(X_r) \end{aligned} \tag{16}$$

5.4.2 Autoencoder

The autoencoder AE trains with the mean squared error (MSE) reconstruction loss function, $L_{AE} = \|x - G(E(x))\|^2$, where $G(E(x))$ is the output of AE and G and E are the decoder (generator) and the encoder of AE. We use weights sharing for the decoder of AE and the generator of CGAN. In this way, we are using both training signals from GAN and AE to train the generator.

5.4.3 Latent Space Discriminator

We further apply a discriminator D_z on top of the encoded space of encoder, $E(x)$, and the random noise, z. The adversarial loss L_{dz} (Eq. 17) forces the encoder to encode images within the distribution of random noise. In this way, we decrease the instability of GAN by keeping its two input representations close to each other. This regularizer helps the model to better discriminate normal and anomalous samples (See Section 5.5.4).

$$L_{dz} = \mathbb{E}_{z \sim P_z}[\log D_z(z, z)] + \mathbb{E}_{z \sim P_z, x \in X}[1 - \log D_z(z, E(x))] \tag{17}$$

5.4.4 Normality Score

AD-CGAN relies on the reconstruction loss of each sample to find anomalies during inference. To see how far a generated image is from the actual test image, we present a normality score, a combination of multiple components. A well-trained AD-CGAN should only be able to generate samples belonging to the normal distribution seen during training. Hence, the normality score, which is defined based on the reconstruction loss, should be lower for normal samples and higher for anomalous samples. Our normality score contains two reconstruction losses: the generation reconstruction loss L_{Gr}, which involves the scores obtained from the trained $CGAN$, and the feature reconstruction loss L_{Fr}, which incorporates the scores obtained from latent representations of a given image. The normality score $NS(x)$ for a given image $x \in \mathcal{D}_{test}$ is defined as the summation of these two losses (Eq. 18).

$$NS(x) = L_{Gr} + L_{Fr} \tag{18}$$

where the generation and feature reconstruction losses are defined as

$$L_{Gr} = \lambda L_{GD}(x) + (1 - \lambda) L_{GR}(x),\ L_{Fr} = \rho L_{FE}(x) + (1 - \rho) L_{FL}(z) \tag{19}$$

Here, L_{Gr} includes discrimination loss, $L_{GD}(x) = \sum |f_D(x) - f_D(G(E(x)))|$ with intermediate representation of a given image x from D_{cgan} as $f_D(x)$, and residual loss, $L_{GR}(x) = \sum |x - G(E(x))|$, similar to the losses presented in (Schlegl et al., 2017). It should be noted that $f_D(x)$ refers to the internal representation of image x obtained from the penultimate layer of the D_{cgan}. L_{Fr} contains encoded L_{FE} and latent L_{FL} feature reconstruction losses (Eq. 20).

$$\begin{aligned} L_{FE}(x) &= \sum |E(x) - E(G(E(x)))| \\ L_{FL}(z) &= \|D_z(E(G(z)), z) - D_z(E(x), E(G(z)))\|_1 \end{aligned} \tag{20}$$

where $E(x)$ is the encoded representation of x from the encoder E, z is the input random noise to the generator G of CGAN, and $G(z)$ is the output of G.

5.5 Experimental Results

We perform extensive experiments on several benchmark image datasets to evaluate our method. The detailed hyperparameters of AD-CGAN are shown in Table 5.1.

5.5.1 Datasets

We considered four benchmark datasets in our experiments: CIFAR-10 (Krizhevsky, 2009), FashionMNIST (fMNIST) (Xiao, Rasul, & Vollgraf, 2017), MNIST (LeCun et al., 1998), and

Table 5.1: AD-CGAN architecture and hyperparameters for the experiments on all datasets. Batch-size of 32 for all, $m = 2, 2, 3, 2$, $n = 4, 4, 4, 5$, $j = 4, 4, 6, 6$, and $k = 2, 2, 2, 2$ for MNIST, FashionMNIST, CIFAR10 and CatsVsDogs dataset respectively.

AD-CGAN architecture

Module	#Layers	Activation fn.	Dropout
$G(z)$	**m** $\times$ $Conv2d$, **n** $\times$ $Trans.Conv2d$	$PReLU$	$\times$
$D_{cgan}(x)$	**j** $\times$ $Conv2d$, **k** $\times$ $Linear$	$LeakyReLU$	0.2
$E(x)$	$5 \times Conv2d$	$LeakyReLU$	0.2
$L_z(z, z)$	$3 \times Linear$	$LeakyReLU$	$\times$

AD-CGAN hyper-parameters

Latent dimension	100 (except CatsVsDogs with 200)
Learning rate	Lr_{CGAN}: 3×10^{-4}, Lr_{AE}: 2×10^{-4}
Optimizer	Adam($\beta_1 = 0.5$, $\beta_2 = 0.999$)
L_{cgan}	$\alpha = 0.3$, $\beta = 0.3$
$NS(x)$	$\lambda = 0.1$, $\rho = 0.5$

CatsVsDogs (Elson, Douceur, Howell, & Saul, 2007). All of the datasets except CatsVsDogs include 10 classes. In order to evaluate AD-CGAN on AD tasks, we employ two different schemes. We introduce soft and hard anomaly detection experiments. In the soft experiments, we consider one-vs-all scheme. In this scheme, a dataset with C classes will lead to C different anomaly detection experiments. A given class c_{ind}, $1 \leq c_{ind} \leq C$, is considered as the *normal* class, while c_{ood} defines *anomalous* class of the rest of $C - 1$ classes. We introduced the hard scheme mainly to show how anomaly detection models based on GANs fail when the inlier class includes multiple distributions. In the hard AD scheme, we introduce all-vs-one scheme. Similar to the soft scheme, each dataset with C classes will lead to C different experiments. However, in contrast with the soft scheme, $1 \leq c_{ood} \leq C$ includes only a single class while c_{ind} contains the remaining $C - 1$ classes. Considering that CatsVsDogs has only two classes, each class was treated as *normal* in a separate experiment.

5.5.2 Baseline Methods

We compare the performance of our model with multiple AD models. DAGMM (Zong et al., 2018), OC-SVM (Schölkopf et al., 1999), TIAE (Cheng, Zhu, Wang, Zhang, & Li, 2021), ALAD (Zenati, Romain, et al., 2018), ADGAN (Deecke et al., 2018), and AE-GAN (Rafiee & Fevens, 2020) are the baselines where the last three models are based on GANs. DAGMM is an autoencoder-based model, which generates a low-dimensional representation of the training data, and leverages a Gaussian mixture model to perform density estimation on the low-dimensional representations. OC-SVM is a kernel-based method that typically uses an RBF kernel to learn a collection of closed sets in the input space. Samples that fall outside of these sets are assumed to

Table 5.2: ROC-AUC (%) comparison of AD models with *one-vs-all* scheme. The symbol † represents results reported from our implementations. All of the results from our implementations are averaged over three different runs.

Datasets	DAGMM (Zong et al., 2018)	OC-CVM (Schölkopf et al., 1999)	ALAD (Zenati, Romain, et al., 2018)	AE-GAN (Rafiee & Fevens, 2020)	ADGAN (Deecke et al., 2018)	TIAE (Cheng et al., 2021)	AD-CGAN
CIFAR-10	$58.7^\dagger$	$62.0^\dagger$	60.7	61.1	60.6	71.2± 1.44	**86.0**± 0.04
fMNIST	$51.8^\dagger$	$92.8^\dagger$	$78.1\pm 0.12^\dagger$	$69.0\pm 0.16^\dagger$	$75.4^\dagger$	86.8± 0.55	**93.9**± 0.02
MNIST	$50.4^\dagger$	$91.7^\dagger$	$62.4\pm 0.09^\dagger$	69.3	91.5	85.2± 0.81	**92.3**± 0.03
CatsVsDogs	$50.6^\dagger$	$51.6^\dagger$	$53.4^\dagger$	$51.6^\dagger$	$49.0^\dagger$	$51.4^\dagger$	**89.8**± 0.04

be anomalous. TIAE uses a transformation invariant autoencoder with an additional training signal based on the most confident inlier samples to find anomalies. ALAD trains a modified bidirectional GAN (Donahue et al., 2017) with multiple discriminators on normal samples and uses an L_1 reconstruction error on the feature space to find anomalies. AE-GAN trains on a mixed model of GAN and autoencoder and uses several scoring components to separate normal and anomalous samples. Unlike the former GAN-based approaches which benefit from a fast inference procedure, ADGAN uses gradient descent to find an inverse mapping of an image to a low-dimensional seed with a GAN trained on normal samples to generate a sample, which makes its inference very slow. ADGAN later uses an L_2 distance between the generated image and the original image to target anomalies.

In this work, we aim to address the difficulties of anomaly detection models based on GANs via introducing a contrastive GAN. Therefore, apart from the comparison with other anomaly detection models, we mainly focus on those AD models which use GANs in their approach. For each of the experiments, if available, we reported their results from their original papers. For AD baselines based on GANs, we also ran their models on all of the datasets within the hard experiments. It should be noted that, due to the long inference process of ADGAN, we ran it only once using their implementation.

5.5.3 Results

The performance of AD-CGAN is summarized in Table 5.2. The Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) measures the performance of a classifier under various threshold settings. In the context of this study, the ROC-AUC is a measurement of how well the classifier can distinguish between normal and anomalous samples. As illustrated in Table 5.2, AD-CGAN outperforms all the baselines in terms of ROC-AUC. The improvement is more notable on CatsVsDogs, with a large improvement for AD-CGAN, and CIFAR-10 by a minimum of 15% improvement on the soft scheme.

The detailed performance of each of the GAN-based models, in soft and hard schemes, for each of the classes of c_{ind}/c_{ood} is presented in Table 5.3. As the table shows, AD-CGAN surpasses all the baselines in each of the individual classes of c_{ind}/c_{ood} except c_1 on the MNIST dataset for both soft and hard schemes. We argue that AD-CGAN performs consistently in all classes of MNIST within both soft and hard schemes, while the performance of ADGAN in the hard scheme is not consistent across classes and their higher performance on c_1 could be related to the specific pattern of this class. This is in contrast with their competitive performance to AD-CGAN in the soft scheme on the MNIST dataset.

Our contrastive GAN without training on any pretext tasks was able to improve the current reconstruction-based anomaly detection models' performance by at least 7% improvement in several experiments. As expected, the performance in the hard scheme is lower compared with the soft scheme since the normal class contains multiple labels, each having a different distribution. This is more notable in FashionMNIST and MNIST datasets with around 7% drop in the performance. We argue that given the similar pattern in several labels of these two datasets, even AD-CGAN with its discriminative representations obtained by the contrastive loss may have difficulty in the hard scheme.

5.5.4 Ablation Study

AD-CGAN is comprised of several training components as well as multiple normality score components. Each of the training components is critical in the models' performance. This can be inferred by comparing the performance of AD-CGAN with each of AE-GAN (Rafiee & Fevens, 2020) and ALAD (Zenati, Romain, et al., 2018) where adding contrastive learning to GAN showed a notable performance gain on the anomaly detection performance. We also argue that the autoencoder is a key element in AD-CGAN where it removes the extensive and time-consuming inference procedure (as stated in the experiments in (Rafiee & Fevens, 2020)). In order to have a better understanding of the effect of each of the components in the proposed normality score, we measure their effects in different anomaly detection settings (see Table 5.4).

Table 5.3: ROC-AUC (%) comparison of GAN-based models on all four datasets with *one-vs-all* and *all-vs-one* schemes. In the *one-vs-all* scheme, the class number defines c_{ind}, while in *all-vs-one*, it refers to c_{ood}. The results are averaged over three different runs. $\lambda = 0.1$ and $\rho = 0.5$ are used for all the experiments. The symbol * represents results reported from the original paper, and it includes the average over the individual classes as well as each individual class. For simplicity, for each of the classes of CIFAR-10 and fMNIST, we use ordinal numbers instead of their label.

		all-vs-one				one-vs-all			
Datasets	Class	ALAD	ADGAN	AE-GAN	Ours	ALAD	ADGAN	AE-GAN	Ours
CIFAR-10	0	61.2± 0.02	44.3	63	**89.8**± 0.12	67	62.7	67	**83.8**± 0.04
	1	61.1± 0.02	39.6	63	**89.5**± 0.13	46	54.6	49	**87.2**± 0.03
	2	40.7± 0.00	58.2	60	**84.7**± 0.06	64	56.1	63	**80.1**± 0.10
	3	48.8± 0.01	44.7	54	**78.6**± 0.09	63	59.5	56	**86.0**± 0.07
	4	35.5± 0.01	66.1	35	**81.7**± 0.14	66	58.6	73	**85.4**± 0.04
	5	53.5± 0.02	44.5	52	**72.8**± 0.01	53	62.8	52	**81.6**± 0.02
	6	47.8± 0.01	61.5	60	**87.6**± 0.06	78	60.4	72	**94.6**± 0.03
	7	49.7± 0.01	47.4	51	**90.7**± 0.03	52	62.3	63	**81.7**± 0.04
	8	52.9± 0.03	45.7	54	**82.6**± 0.02	75	70.2	68	**89.3**± 0.06
	9	59.4± 0.01	31.3	63	**86.9**± 0.02	43	59.1	48	**90.7**± 0.04
	Average	51.1± 0.08	48.3	55.5*	**84.5**± 0.05	60.7*	60.6*	61.1*	**86.0**± 0.04
fMNIST	0	54.0± 0.02	48.4	45.3± 0.09	**89.5**± 0.06	79.4± 0.02	74.1	74.4± 0.03	**93.3**± 0.04
	1	68.2± 0.04	63.7	32.8± 0.12	**85.8**± 0.03	94.1± 0.04	92.3	92.3± 0.01	**95.9**± 0.04
	2	55.5± 0.03	40.4	57.9± 0.08	**90.4**± 0.07	60.6± 0.09	71.1	67.7± 0.03	**94.0**± 0.06
	3	47.9± 0.04	60.5	23.0± 0.05	**81.7**± 0.10	79.5± 0.05	81.6	80.0± 0.03	**93.7**± 0.03
	4	60.3± 0.13	47.8	34.9± 0.07	**81.4**± 0.02	76.4± 0.06	73.6	82.5± 0.01	**93.1**± 0.02
	5	22.2± 0.01	66.9	80.4± 0.02	**93.8**± 0.04	85.5± 0.01	77.3	36.5± 0.06	**93.5**± 0.02
	6	45.1± 0.01	34.5	52.1± 0.06	**92.1**± 0.02	61.2± 0.08	70.0	55.1± 0.04	**91.7**± 0.03
	7	44.1± 0.05	67.1	55.5± 0.09	**87.0**± 0.09	94.9± 0.02	91.0	77.9± 0.07	**98.5**± 0.01
	8	50.2± 0.09	54.1	76.0± 0.02	**83.8**± 0.01	62.6± 0.03	50.3	49.9± 0.06	**91.6**± 0.08
	9	60.7± 0.07	56.6	63.6± 0.09	**91.1**± 0.06	86.5± 0.13	73.2	73.4± 0.13	**93.7**± 0.07
	Average	50.8± 0.12	54.0	52.2± 0.18	**87.7**± 0.04	78.1± 0.12	75.4	69.0± 0.16	**93.9**± 0.02
MNIST	0	61.0± 0.05	42.7	73	**94.6**± 0.07	74.7± 0.12	97.2	85	**97.1**± 0.02
	1	87.1± 0.03	**93.1**	56	85.8± 0.11	69.8± 0.16	**99.7**	98	95.7± 0.02
	2	44.5± 0.04	39.7	61	**91.6**± 0.05	50.4± 0.09	87.4	54	**92.1**± 0.03
	3	47.7± 0.05	61.2	55	**86.6**± 0.08	65.7± 0.05	84.8	69	**91.2**± 0.03
	4	56.7± 0.05	70.2	49	**77.4**± 0.04	63.6± 0.06	91.0	72	**95.5**± 0.01
	5	50.1± 0.06	53.1	49	**82.8**± 0.03	56.1± 0.04	**91.6**	54	87.8± 0.02
	6	51.8± 0.11	59.8	55	**86.9**± 0.08	53.0± 0.08	**95.7**	60	88.6± 0.06
	7	56.4± 0.09	75.2	44	**76.1**± 0.02	49.6± 0.01	**93.7**	68	92.5± 0.01
	8	41.2± 0.08	58.5	59	**83.9**± 0.02	75.3± 0.07	81.6	69	**87.2**± 0.05
	9	42.4± 0.02	71.1	56	**84.5**± 0.06	65.2± 0.09	92.4	64	**95.2**± 0.02
	Average	53.9± 0.13	62.5	55.7*	**85.0**± 0.05	62.4± 0.09	91.5*	69.3*	**92.3**± 0.03
CatsVsDogs	Cats	-	-	-	-	52.6	53.1	51.7	**92.7**± 0.03
	Dogs	-	-	-	-	54.1	44.9	52.1	**86.9**± 0.05
	Average	-	-	-	-	53.4	49.0	51.6	**89.8**± 0.04

Feature reconstruction loss is added to the normality score to measure how discriminative the latent representations of the two reconstruction models are. Several experiments on MNIST and FashionMNIST on soft and hard AD schemes showed that adding D_z leads to more discriminative latent representation, which affects the normality scores obtained by the feature reconstruction loss. We defined four distinct models of AD-CGAN based on the normality score components they

Table 5.4: Ablation studies on MNIST and FashionMNIST given different normality score components of AD-CGAN. We used $\lambda = 0.1$ where the generation reconstruction loss had been used. We set $\rho = 1$ and $\rho = 0$ for AD-CGAN$_{LFE}$ and AD-CGAN$_{LFL}$, respectively. The ROC-AUC (%) results are averaged over three different runs.

	MNIST (%)		fMNIST (%)	
model	all-vs-one	one-vs-all	all-vs-one	one-vs-all
$AD-CGAN_{LG}$	56.8± 0.13	72.3± 0.12	68.7± 0.11	80.8± 0.11
$AD-CGAN_{LFE}$	66.6± 0.12	84.4± 0.11	70.4± 0.13	86.1± 0.08
$AD-CGAN_{LFL}$	73.2± 0.09	84.9± 0.05	74.5± 0.08	87.0± 0.05
$AD-CGAN_{GF}$	**85.0**± 0.05	**92.3**± 0.03	**87.7**± 0.04	**93.9**± 0.02

have access to: AD-CGAN$_{LG}$ represents AD-CGAN with only generation reconstruction loss; AD-CGAN$_{LFL}$ with only latent feature reconstruction loss, L_{FL}; AD-CGAN$_{LFE}$ with only encoded feature reconstruction loss, L_{FE}; and AD-CGAN$_{GF}$ contains both L_{Gr} and L_{Fr} in its normality score. It should be mentioned that in each of these models, generation reconstruction loss is considered as part of the normality score. As the results reveal, removing the feature reconstruction loss (ignoring D_z) negatively affects the performance of AD-CGAN. The impact is more severe in the case of the all-vs-one (hard) scheme. On the other hand, AD-CGAN$_{LFE}$ that trains with D_z and encoded feature reconstruction loss, significantly improved AD-CGAN$_{LG}$ in both datasets. Similar behavior is observed on the results on AD-CGAN$_{LFL}$. However, it is important to note that in all the experiments, ignoring any training and/or normality score component results in lower performance. The best results are achieved when all the components with the right amount of contributions are considered, as it is shown in AD-CGAN$_{GF}$.

To further validate the effect of the contrastive loss, in another experiment, we found that applying contrastive loss to ADGAN (Deecke et al., 2018) improves the ROC-AUC by 3% and 9% on CIFAR10 and FashionMNIST on all-vs-one, respectively.

5.6 Conclusion and Future Work

We presented a new reconstruction-based approach to tackle the problem of anomaly detection (AD) in images. The proposed approach adds contrastive learning to an anomaly detection model based on a generative adversarial network (GAN), AD-CGAN, to learn more discriminative and task agnostic features of normal data. AD-CGAN uses a normality score function including multiple components to further separate normal and anomalous samples. In this study, we considered two different AD schemes, soft and hard, to evaluate the performance of AD-CGAN. AD-CGAN was able to outperform all the previously reconstruction-based approaches on all four benchmark datasets within both soft and hard schemes. These results may open a new path for reconstruction-based anomaly detection models leading to more discriminative representations of normal data.

Chapter 6

Source-free Domain Adaptation Requires Penalized Diversity

This chapter focuses on mitigating distribution shifts rather than detecting them. In this work, assuming covariate and label distribution shifts are present, we study knowledge transfer between different domains in the absence of source data. Since diversity in representation space can be vital to a model's adaptability in varied and difficult domains, we study the effect of increasing diversity in an ensemble alongside a weighted regularizer to tackle covariate and label distribution shifts simultaneously.

6.1 Introduction

In recent years, the field of machine learning (ML) has witnessed immense progress in computer vision (He et al., 2016), natural language processing (Vaswani et al., 2017), and speech recognition (Bahdanau, Chorowski, Serdyuk, Brakel, & Bengio, 2016) due to the advances of deep neural networks (DNNs). Despite the increasing popularity of DNNs, they often perform poorly on unseen distributions (Geirhos et al., 2020), leading to overconfident and miscalibrated models. Combining the predictions of several models seems to be a feasible way to improve the generalizability of these models (Turner & Oza, 1999). On account of its simplicity and effectiveness, ensemble learning became popular in many machine learning applications. Due to the i.i.d. assumption that training and test sets are drawn from the same distribution, calibration (Dawid, 1982) is introduced to the traditional machine learning paradigm to elucidate the model uncertainty. Additionally, predictive uncertainty is crucial under dataset shift–when confronted with a sample from a shifted distribution, an ideal model should reflect increased uncertainty in its prediction.

Commonly, a dataset distribution shift can occur due to diverse sources (Quiñonero-Candela et al., 2008): (i) domain shift, also known as covariate shift, is caused by hardware differences in data acquisition devices; (ii) feature distribution disparity is caused by population-level differences (e.g., gender, ethnicity) across domains; (iii) label distribution shift, where the proportional prevalence of labels in the source domain differs from that of the target domain. Due to the variety of distribution shifts, models have failed in real-world applications with shifted domains, thus posing an important threat to safety-critical applications.

Hypothesis transfer learning (HTL), also referred to as source-free domain adaptation (SFDA), addresses distribution shift under the non-transductive setting by using knowledge encoded in a model pretrained on the source domain to inform learning on the target domain. Unlike traditional domain adaptation (DA) approaches, SFDA models do not have simultaneous access to the data from both source and target domains. This assumption mitigates the privacy and storage concerns arising in conventional DA methods.

Extending ensemble learning to DA frameworks and, in particular, SFDA methods can uncover multiple modes within the source domain, improving the transferability of these models (Lao et al., 2021). However, the performance gain of an ensemble model is largely related to the diversity of its members. Particularly, averaging over identical networks or ensemble members with limited diversity is not better than a single model (Rame & Cord, 2021).

In this work, we encourage diversity among ensemble members in an unsupervised source-free domain adaptation setting where no labeled target data is available. While recent work in unsupervised SFDA has shown promising results, it either relies on a unique feature extractor (Liang et al., 2020), or one shared between an ensemble of source hypotheses (Lao et al., 2021), which leads to limited diversity in the function space of the source domain (see Sec. 6.4.5 for analysis).

Diversity in ensemble leads to the best-calibrated uncertainty estimators (Lakshminarayanan et al., 2017), and therefore the performance benefits of feature diversity within ensembles in out-of-distribution (OOD) settings (Pagliardini et al., 2022). Other recent works in DNN analysis also show that different architectures tend to explore different representations (Antorán et al., 2020; Kornblith, Norouzi, Lee, & Hinton, 2019; Nguyen, Raghu, & Kornblith, 2021; Zaidi et al., 2021). Inspired by them , our work proposes to increase diversity by not only using separate feature extractors but also by introducing Distinct Backbone Architectures (DBA) across hypotheses.

While a regularization approach to unconstrained mutual information (MI) maximization during adaptation is promising in low diversity settings (Lao et al., 2021), enforcing similarity between highly diverse hypotheses is insufficient to counteract the catastrophic impact of weak hypotheses when they inevitably arise as outliers. Therefore, we highlight the necessity of a trade-off between diversity and the amount of freedom each ensemble member can have. Hence, we introduce Penalized Diversity (PD), a new unsupervised SFDA approach that maximizes diversity exploitation via DBA while mitigating the negative impact of Weak Hypotheses through the Penalization (WHP) of their contribution by regularization.

In many real-world applications, the uniform distribution assumption between source and target does not hold. This assumption can negatively impact the performance of many current SFDA models under label distribution shift (Lao et al., 2021; Liang et al., 2020) (Sec. 6.4.4, label shift experiments). We further extend PD to address the label distribution shift by introducing a weighted MI maximization based on estimation over target distribution. Extensive experiments on multiple domain adaptation benchmarks (Office-31, Office-Home, and VisDA-C), medical, and digit datasets under covariate and label distribution shifts exhibit the effectiveness of PD.

6.2 Related Work

Transfer learning approaches can be divided into data-driven and model-driven approaches. Data-driven approaches such as instance weighting (Bickel et al., 2007; Zadrozny, 2004) and domain adaptation models such as DAN (Long et al., 2015), DANN (Ganin & Lempitsky, 2015) and MDD (Y. Zhang et al., 2019) assume to have direct access to the source data during the knowledge transfer. To mitigate transfer learning models' privacy and storage concerns, source-free domain adaptation (SFDA) approaches (also known as model-driven or hypothesis transfer learning) are proposed. SFDA is a transfer learning strategy where a model trained on the source domain incorporates the learning procedure of the target domain. It was first introduced by Kuzborskij and Orabona (2013) where the access to the source domain was only limited to a set of hypotheses induced from it, unlike domain adaptation, where both source and target domains are used to adapt the source hypothesis to the target domain.

Source-free Domain Adaptation has been investigated from both practical and theoretical points of view in computer vision applications. Several studies have analyzed the effectiveness of SFDA on various specific ML algorithms (Kuzborskij & Orabona, 2013, 2017; X. Wang & Schneider, 2015), while others proposed more generally applicable frameworks (Du, Koushik, Singh, & Póczos, 2017; Fernandes & Cardoso, 2019). These studies can be divided based on the availability of labeled data in the target domain. Most previous studies considered the supervised SFDA setting (labeled target domain) (Du et al., 2017; Fernandes & Cardoso, 2019; Kuzborskij & Orabona, 2013, 2017; X. Wang & Schneider, 2015), while unsupervised SFDA (uSFDA) (unlabeled target domain) has only recently gained interest (Lao et al., 2021; Liang et al., 2020). SFDA models mostly rely on a single hypothesis to transfer knowledge to the target domain. Lao et al. (2021) showed that using a single hypothesis for uSFDA is prone to overfitting the target domain and causes catastrophic forgetting of the source domain. They were the first to propose using multiple hypotheses to mitigate this effect. More recently (F. Wang, Han, Gong, & Yin, 2022) proposed a novel way to tackle the SFDA problem by finding domain-invariant parameters rather than domain-invariant features in the model.

Ensemble Models Recently, deep neural network calibration gained considerable attention in the machine learning research community. Previous studies explored the effect of Monte Carlo dropout (Gal & Ghahramani, 2016; Kingma et al., 2015) and variational inference methods (Maddox, Izmailov, Garipov, Vetrov, & Wilson, 2019). However, it has been shown that the best-calibrated uncertainty estimators can be achieved by neural network ensembles (Ashukha, Lyzhov, Molchanov, & Vetrov, 2020; Lakshminarayanan et al., 2017; Ovadia et al., 2019). The importance of well-calibrated models becomes more important under the presence of dataset shift. The success of ensemble models is mainly related to the diversities present between their members. Ensemble diversity has been widely investigated in the literature (Brown, 2004; Y. Liu & Yao, 1999). Improving diversity in neural network ensembles has become a focus in recent work. Stickland and Murray (2020) suggest augmenting each member of an ensemble with a different set of augmented inputs to increase the diversity among members. While a few recent studies propose deep ensemble models based on different neural network architectures to ensure diversity (Antorán et al., 2020; Zaidi et al., 2020).

Recently Pagliardini et al. (2022) suggest that encouraging diversity between the ensemble predictions helps to generalize in the OOD setting by increasing disagreements and uncertainties over out-of-distribution samples. Whereas Y. Lee et al. (2022) introduced an ensemble of multiple hypotheses with shared feature extractors and separate classifier heads to generalize in the presence of spurious features. They proposed to increase diversity among the classifiers through mutual information minimization over the hypotheses predictions on unlabeled target data. Our work is different from (Y. Lee et al., 2022; Pagliardini et al., 2022) in a sense that (i) to increase diversity, PD does not need a carefully selected set of target samples unlike both (Y. Lee et al., 2022; Pagliardini et al., 2022), (ii) different from (Pagliardini et al., 2022) that limits the model to have a smaller or equal number of hypotheses than the total number of classes, we have freedom over the number of hypotheses in our model, and (iii) WHP mitigate weak hypotheses to improve overall performance without requiring access to labeled target samples as opposed to the active query strategy presented in (Y. Lee et al., 2022). Despite its performance, PD also has its own limitation. Its diversification and penalization approaches force PD to be more effective with an ensemble of at least three hypotheses.

Label Distribution Shift Many domain adaptation studies focus only on covariate shift. Despite the negative impact of label distribution shift in transferring knowledge, it has been mostly neglected. Learning domain-invariant representations and using estimated class ratios between domains as importance weights in the training loss became a dominant strategy for many recent practices (Gong et al., 2016; Shui et al., 2021; Tachet des Combes, Zhao, Wang, & Gordon, 2020). Rakotomamonjy et al. (2022) proposed MARS to learn domain-invariant representations with sample re-weighting. Several studies attempt to benefit from optimal transport (OT) to find a transport function from source to target with a minimum cost. Kirchmeyer et al. (2022) proposed a reweighing model which maps pretrained representations using OT. The key difference between these models and our modified MI solution is that they all assume accessing the source data during adaptation and their reweighing strategies are mainly based on source and target ratios.

6.3 Approach

6.3.1 Preliminaries

Assuming $\mathcal{X}$ indicates the input space and $\mathcal{Y}$ represents the output space, in an unsupervised source-free domain adaptation (uSFDA) setting, we denote the source domain as $\mathcal{D}_s = \{(x_i^S, y_i^S)\}_{i=1}^{N_s}$, where $x_i^S \in \mathcal{X}^\mathcal{S}$ and $y_i^S \in \mathcal{Y}^\mathcal{S}$. The unlabeled target domain is denoted as $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$, where $x_i^T \in \mathcal{X}^\mathcal{T}$ and $\mathcal{X}^S \neq \mathcal{X}^T$. For now, we assume that the difference in the joint distribution $P(\mathcal{X}, \mathcal{Y})$ of source and target stems from the covariate shift only. Therefore, this induces a domain shift between the source and target domains, ($P_S(X) \neq P_T(X)$), whereas the learning task remains the same, with $\mathcal{Y}^\mathcal{S} = \mathcal{Y}^\mathcal{T}$ and $P_S(Y \mid X) = P_T(Y \mid X)$. Given a hypothesis space $\mathcal{H}$, uSFDA learns a source hypothesis $h_s : \mathcal{X}^S \rightarrow \mathcal{Y}^S \in \mathcal{H}^S$ and a target hypothesis $h_t : \mathcal{X}^T \rightarrow \mathcal{Y}^T \in \mathcal{H}^T$, to predict the unobserved target labels Y_t^*. From the Bayesian perspective, the predictive posterior distribution can be written as:

$$p(Y_t^* \mid \mathcal{D}_s, \mathcal{D}_t) = \int_{h_t} p(Y_t^* \mid \mathcal{D}_t, h_t) \int_{h_s} p(h_t \mid \mathcal{D}_t, h_s) p(h_s \mid \mathcal{D}_s) dh_s dh_t \tag{21}$$

Eq. 21 describes two learning phases; first, the posterior over the source hypothesis $p(h_s \mid \mathcal{D}_s)$ is learned using the source dataset $\mathcal{D}_s$, and second, the posterior over the target hypothesis $p(h_t \mid \mathcal{D}_t, h_s)$ is learned by marginalizing over samples of the source hypothesis adapted to the target domain, which only contains unlabeled examples.

Liang et al. (2020) use a single model to estimate the distribution over the source hypothesis, and by extension the distribution over the target hypothesis. Lao et al. (2021) improved this approximation by incorporating multiple hypotheses that *share* the same feature extraction backbone. While the latter is considered an ensemble, by definition, it is constrained by learning shared extracted features. In this paper, we promote diversity by introducing the use of Distinct Backbone

Architectures (DBA) across hypotheses. We argue and show empirically (Sec. 6.4.4) that this helps us achieve a better approximation of $p(h_t \mid \mathcal{D}_t, h_s)$ with higher diversity in the representation space.

However, unconstrained MI maximization during adaptation is prone to the induction of weak hypotheses due to error accumulation. The hypothesis disparity (HD) introduced by (Lao et al., 2021) acts as a regularizer by enforcing similarity across hypotheses over the distribution of predicted labels. While this regularization showed promise in low diversity settings, enforcing similarity between highly diverse hypotheses is insufficient, and weak hypotheses inevitably arise (see Sec. 6.4.5 for experiments). Unfortunately, the very nature of how similarity is computed in HD makes it highly vulnerable to weak hypotheses. We propose an approach that mitigates the negative impact of Weak Hypotheses through the Penalization (WHP) of their contribution to the computation of HD when they arise as outliers (Sec. 6.4.4).

In the following sections, we describe three main components of our proposed model, Penalized Diversity (PD).

6.3.2 Learning Diverse Source Hypotheses Using Distinct Backbone Architectures

To maximize the diversity of predictive features learned in the source domain, we propose removing any weight sharing between the backbones of separate hypotheses by introducing the use of distinct architectures. For example, on the LIDC dataset, our approach (DBA) is implemented through the use of a mixture of ResNet10 and ResNet18 backbones.

We define the set of source hypotheses as $\{h_i^S : h_i^S = f_i^S \circ g_i^S\}_{i=1}^M$, where $\{f_i^S\}_{i=1}^M$ and $\{g_i^S \in \Psi^S\}_{i=1}^M$ represent the set of classifiers and the set of feature extractors, respectively, and M represents the number of hypotheses. We train each hypothesis using the cross entropy loss function (CE):

$$L_S = \arg\min_{\mathcal{H}^S} \mathbb{E}_{(x,y)\in\mathcal{X}^S\times\mathcal{Y}^S}[CE(h(x), y)],\ \forall h \in \{h_i^S\}_{i=1}^M \tag{22}$$

where $h(x) = p(y \mid x; h)$ denotes the probability distribution of input x predicted by hypothesis h .

6.3.3 Diversity Exploitation Through Weak Hypothesis Penalization

Assuming $\mathcal{D}_t = \{(x_i^T)\}_{i=1}^{N_t}$ is a set of unlabeled target samples, our goal is to effectively adapt the set of hypotheses trained on the source domain $\{h_i^S\}_{i=1}^{M}$ into a set of target hypotheses $\{h_i^T\}_{i=1}^{M}$. Due to the absence of both source data and labeled target data during the adaptation phase, we maximize the mutual information (MI) between the target data distribution (X^T) and the predictions by the target hypotheses ($\hat{Y}^T$) (Liang et al., 2020) using Eq. 23.

$$\max_{\Psi^T} \mathbb{E}_{x \in \mathcal{X}^T}[I(X^T; h(X^T))], \ \forall h \in \{h_i^T\}_{i=1}^{M} \tag{23}$$

where MI is defined as $I(X^T; \hat{Y}^T) = H(\hat{Y}^T) - H(\hat{Y}^T \mid X^T)$ with H indicating entropy, $\hat{Y}^T$ is the predicted output of $h(X^T)$, and Ψ^T is the space of target feature extractors. Assuming that only covariate shift is present, both the source and the target domains share the same label space, so we keep the parameters for the classifiers f_i^T fixed while updating the feature extractors $g_i^T \in \Psi^T$.

Unconstrained unsupervised training of target hypothesis ensembles solely using MI maximization results in undesirable target label prediction disagreements. We use the hypothesis disparity (HD) regularization to marginalize out these disagreements (Lao et al., 2021). HD measures the dissimilarity between the predicted label probability distributions among pairs of hypotheses over the input space $\mathcal{X}$:

$$\text{HD}_{h_i, h_j \in \{h^T\}, i \neq j}(h_i, h_j) = \int_{\mathcal{X}} d(h_i(x), h_j(x)) p(x) dx \tag{24}$$

where $d(.)$ defines the dissimilarity metric. Throughout this study, we use cross entropy to measure dissimilarity.

In its original formulation, computing HD relies on randomly selecting a single hypothesis that serves as an anchor (reference) for the pairwise disparity measures with the rest of the $M-1$ hypotheses. This selected anchor remains fixed throughout the training process. We note that this method of choosing the anchor may potentially have a catastrophic impact; a weak performance hypothesis chosen as the anchor could act as an attractor and collapse the model (See 6.4.5 and Table 6.7).

In order to address this issue, we redefine HD (Eq. 24) by proposing Weak Hypothesis mitigation through Penalization (WHP), which constructs the anchor hypothesis (h_{whp}) as an ensemble where the contribution of each hypothesis is weighted according to its cosine similarity to other hypotheses (Eq. 25 to Eq. 28):

$$h_{\text{whp}}(X^T) = \sum_{i=1}^{M} \hat{w}_i h_i(X^T), \tag{25}$$

where $h_i \in \{h^T\}$, and $\hat{w}_i$ represents the normalized weight for each hypothesis and is computed as:

$$\hat{w}_i = \frac{\exp(w_i)}{\sum_{j=1}^{M} \exp(w_j)} \tag{26}$$

$$w_i = \mathbb{E}_{h_j \in \{h^T\}, j \neq i} \left[\frac{h_i(X^T) \cdot h_j(X^T)}{|h_i(X^T)| \cdot |h_j(X^T)|} \right], \ \forall i. \tag{27}$$

$$\text{HD}_{h_i \in \{h^T\}}(h_i, h_{\text{whp}}) = \int_{\mathcal{X}} d(h_i(x), h_{\text{whp}}(x)) p(x) dx \tag{28}$$

In effect, the contribution to the ensemble anchor h_{whp} of a more distant hypothesis h_i, based on its marginal cosine similarity to other hypotheses, is penalized through the reduction of w_i. Hence, we improve performance by diminishing the probability of selecting a weak anchor. In Section 6.4.5, we compare WHP to alternative strategies for anchor hypothesis selection and show its superior performance.

6.3.4 Target Distribution Estimation Via Pseudo-Labels

We take one step further and refine our assumption on possible shifts in the joint distribution of source and target. Instead of assuming that the changes only stemmed from the difference in their marginal ($P(X)$), we also allow shifts in their prior, i.e. $P_S(Y) \neq P_T(Y)$, namely label distribution shift. For PD to be able to perform under label distribution shift, we suggested weighted mutual information (MI) based on the proportion of target classes. Since the exact proportion of target classes is not accessible in an unsupervised setting, an estimation using pseudo-labeling (Liang et al., 2020) can be used. The label entropy for MI maximization is reweighted according to the estimated class proportions and reformulates the MI maximization as:

$$I_W(X^T; \hat{Y}^T) = W * H(\hat{Y}^T) - H(\hat{Y}^T \mid X^T) \tag{29}$$

where $W = [w_1, \ldots, w_C]$ is an estimated class proportion from pseudo-labels, and $w_i = \frac{n_{c_i}}{\sum_{j=1}^{C} n_{c_j}}$, where n_{c_i} represents the number of samples in class c_i and C is the total number of classes.

In summary, our full objective for target training is a combination of weighted mutual information and hypothesis disparity regularization.

$$L_T = \alpha \mathbb{E}_{h \in \mathcal{H}^T}[-I_W(X^T; h(X^T))] + \beta \mathbb{E}_{h \in \{h^T\}}[\mathrm{HD}(h, h_{\text{whp}})] \tag{30}$$

where α and β are hyperparameters indicating the contribution of each of MI and HD in the target training.

6.4 Experiments and Results

6.4.1 Datasets

To validate our model under covariate shift, we consider natural and medical image datasets. For the natural images, we consider domain adaptation benchmark datasets, namely **Office-31** (Saenko, Kulis, Fritz, & Darrell, 2010), **Office-Home** (Venkateswara, Eusebio, Chakraborty, & Panchanathan, 2017), and **VisDA-C** (Peng et al., 2018). For the medical application, we evaluate our model on the **LIDC** (Armato III et al., 2011) dataset. Office-31 dataset includes three domains that share a set of 31 classes; Amazon (A), DSLR (D), and Webcam (W). Office-Home has four domains, each having 65 classes; Artistic images (AR), Clip art (CL), Product images (PR), and Real-World images (RW). VisDA-C has 12 classes with synthetic images in the source domain and real images in the target domain. For our medical imaging experiment, we divided the LIDC dataset into four domains based on the manufacturer of the data-capturing device: GE_medical (G), Philips (P), SIEMENS (S), and TOSHIBA (T). Each of these domains has two classes, healthy and unhealthy.

We validate the existence of covariate shift across LIDC domains, from both statistical and experimental perspectives. A detailed statistical analysis is provided in the supplementary material 6.6. For the experiments on label shift, we consider synthetic digit datasets–**MNIST** (M) (LeCun et al., 1998), **MNIST-M** (N) (Ganin et al., 2016), and **USPS** (U) (Hull, 1994).

6.4.2 Baselines

Unsupervised transfer learning approaches can be categorized as either unsupervised domain adaptation (UDA) or SFDA, depending on whether or not they require access to the source data during the adaptation phase. We consider baselines from both sets. For UDA, we compare PD to DANN (Ganin & Lempitsky, 2015), DAN (Long et al., 2015), CDAN (Long, Cao, Wang, & Jordan, 2018), SAFN+ENT (Xu, Li, Yang, & Lin, 2019), rRevGrad+CAT (Z. Deng, Luo, & Zhu, 2019), MDD (Y. Zhang et al., 2019), and MCC (Jin, Wang, Long, & Wang, 2020). For SFDA, we use AdaBN (Li, Wang, Shi, Liu, & Hou, 2016), Tent (D. Wang, Shelhamer, Liu, Olshausen, & Darrell, 2021), SHOT (Liang et al., 2020), HDMI (Lao et al., 2021), and NRC (S. Yang, van de Weijer, Herranz, Jui, et al., 2021). We also consider the performance of source hypotheses at directly predicting target labels as a Source-only model, and MI-ensemble as a model with three hypotheses with only MI maximization and no regularizer.

For label distribution shift experiments, aside from comparing with two SFDA models namely SHOT and HDMI, we compare PD with MARS (Rakotomamonjy et al., 2022) and OSTAR (Kirchmeyer et al., 2022), the two recent state-of-the-art models for label distribution shift. MARS (Rakotomamonjy et al., 2022) proposed based on two estimating proportion strategies, where hierarchical clustering defines MARSc and Gaussian mixtures indicates MARSg. In nearly all our experiments, MARSc outperforms MARSg, therefore we only report the performance of MARSc while referring to it as MARS.

6.4.3 Experimental Setup

In the experiments presented in this paper, we consider both covariate shift and label distribution shift between source and target domains. We simply instill diversity in DBA using different depths of a given architecture (Antorán et al., 2020; Zaidi et al., 2021). The code base of PD is built upon

SHOT.

For the medical dataset, we used different depths of 3D-ResNet as the hypothesis backbone, mainly ResNet10 and ResNet18. Each backbone $\{\phi_i\}_{i=1}^{M}$ is then followed by a set of fully connected layers, Batch-Norm, ReLU activation function, and Dropout referred to as the bottleneck layer. We used 512 as the dimension of extracted features. For the classifier $\{f_i\}_{i=1}^{M}$, we used a shallow neural network with two fully connected layers, followed by a ReLU activation function, and Dropout. Each model trained for 200 iterations with batch size 32 and learning rate $1e-4$ and AdamW optimizer on the source domain. We used the same configuration for the target training except the learning rate was decreased to $1e-5$. For the experiments with 3 hypotheses, we used ResNet $\{10, 18, 10\}$, indicating the depth of each feature extractor. For the baselines, we used the same configuration with ResNet18 as their backbone whether they have single or multiple hypotheses. $\alpha = 0.3$ and $\beta = 0.5$ are used for the target training.

For natural images datasets, we use different depths of ResNet (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) as the backbone of our feature extractors. Specifically, ResNet of depths $\{34, 50, 34\}$ for Office-31 and Office-Home and ResNet of depths $\{50, 101, 110\}$ for VisDA-C are chosen as the depths of $\{g_i\}_{i=1}^{M}$ for $M = 3$ hypotheses. The same bottleneck layer as our experiments on medical data is used for the experiments on natural images. We followed similar hyperparameters as (Liang et al., 2020) for synthetic digit datasets with different depths on PD. For both natural and synthetic datasets, we trained the source hypotheses for 5k iterations, with learning rate $3e-4$ and batch size of 64. Target hypotheses are trained for 20k iterations with learning rate $3e-4$ and batch size of 64. We used SGD optimizer for both source and target training. We used $\alpha = 1$ and $\beta = 0.5$ for the target training objective function.

For the experiments on digit datasets, we used $p = 0.1$ as the probability value of changing the chosen class, i.e. a class with 1000 samples in the target domain will be reduced to 100 samples in the new shifted domain.

6.4.4 Results

In this section, we present the performance of our model in comparison with the baselines in each benchmark dataset under different distributional shifts.

Table 6.1: Target accuracy (%) on Office-31 under covariate shift (source → target). In this and all the following tables, † represents results reported from our implementations.

Method	Source-free	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Source-only†	×	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DAN (Long et al., 2015)	×	78.6	80.5	63.6	97.1	62.8	99.6	80.4
DANN (Ganin & Lempitsky, 2015)	×	79.7	82.0	68.2	96.9	67.4	99.1	82.2
SAFN+ENT (Xu et al., 2019)	×	90.7	90.1	73.0	98.6	70.2	99.8	87.1
rRevGrad+CAT (Z. Deng et al., 2019)		90.8	94.4	72.2	98.0	70.2	**100.**	87.6
MDD (Y. Zhang et al., 2019)	×	93.5	**94.5**	74.6	98.4	72.2	**100.**	88.9
MCC (Jin et al., 2020)	×	95.5	98.6	100	94.4	72.9	74.9	89.4
MI-ensemble†	✓	91.0	93.0	72.3	96.5	73.7	97.4	87.3
AdaBN (Li et al., 2016)	✓	81.0	82.4	67.2	97.7	68.2	99.8	82.7
Tent (D. Wang et al., 2021)	✓	82.1	85.1	68.8	97.5	63.0	99.8	82.7
SHOT (Liang et al., 2020)	✓	93.1	90.9	74.5	98.8	74.8	99.9	88.7
HDMI (Lao et al., 2021)	✓	94.4	94.0	73.7	98.9	75.9	99.8	89.5
NRC (S. Yang et al., 2021)	✓	**96.0**	90.8	**75.3**	**99.0**	75.0	**100.**	89.4
PD	✓	95.6	94.3	**75.3**	98.7	**76.4**	99.8	**90.0**

Table 6.2: Target accuracy (%) on Office-Home under covariate shift (source → target).

Method	Source-free	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
Source-only†	×	45.6	69.2	76.5	55.3	64.4	67.4	55.1	41.6	74.4	66.0	46.3	79.4	61.8
DAN (Long et al., 2015)	×	43.6	57.0	67.9	45.8	56.5	60.4	44.0	43.6	67.7	63.1	51.5	74.3	56.3
DANN (Ganin & Lempitsky, 2015)	×	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
SAFN (Xu et al., 2019)	×	52.0	71.7	76.3	64.2	69.9	71.9	63.7	51.4	77.1	70.9	57.1	81.5	67.3
MDD (Y. Zhang et al., 2019)	×	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	**60.2**	82.3	68.1
MI-ensemble†	✓	55.2	71.9	80.2	62.6	76.8	77.8	63.2	53.8	81.1	67.9	58.3	81.4	69.2
AdaBN (Li et al., 2016)	✓	50.9	63.1	72.3	53.2	62.0	63.4	52.2	49.8	71.5	66.1	56.1	77.1	61.5
Tent (D. Wang et al., 2021)	✓	47.9	66.0	73.3	58.8	65.9	68.1	60.2	47.3	75.4	70.8	54.0	78.7	63.9
SHOT (Liang et al., 2020)	✓	56.9	78.1	81.0	67.9	78.4	78.1	67.0	54.6	81.8	73.4	58.1	**84.5**	71.6
HDMI (Lao et al., 2021)	✓	57.8	76.7	81.9	67.1	78.8	**78.8**	66.6	55.5	82.4	73.6	59.7	84.0	71.9
NRC (S. Yang et al., 2021)	✓	57.7	**80.3**	**82.0**	68.1	**79.8**	78.6	65.3	56.4	83.0	71.0	58.6	85.6	72.2
PD	✓	**58.9**	78.0	80.9	**69.3**	76.7	76.9	**69.6**	**56.5**	**83.4**	**75.1**	59.9	**84.5**	**72.5**

Table 6.3: Target accuracy (%) on LIDC under covariate shift (source → target). The results are averaged over five different runs.

Method	Source-free	G→P	G→S	G→T	P→G	P→S	P→T	S→G	S→P	S→T	T→G	T→P	T→S	Avg.
Source-only†	×	65.6	65.9	45.9	61.9	60.5	49.3	65.2	66.9	57.1	50.0	50.0	50.0	57.2
DAN (Long et al., 2015)†	×	59.3	65.8	33.9	61.1	59.7	30.7	56.9	58.6	40.2	48.7	45.9	47.6	50.7
MDD (Y. Zhang et al., 2019)†	×	64.1	63.6	57.7	54.7	59.8	47.7	**67.2**	63.4	59.1	50.0	50.3	50.0	57.3
MI-ensemble†	✓	67.1	63.4	**66.6**	62.9	63.2	52.1	65.1	64.9	55.2	60.8	58.6	58.2	61.5
SHOT (Liang et al., 2020)†	✓	67.0	**67.1**	61.6	59.9	56.9	53.2	66.8	**69.0**	66.1	60.4	61.0	54.9	61.9
HDMI (Lao et al., 2021)†	✓	67.1	66.6	64.6	65.4	64.2	54.6	66.2	65.1	54.6	60.7	**60.0**	58.7	62.3
PD	✓	**68.9**	65.9	60.9	**65.6**	**65.6**	**54.8**	65.8	66.9	**66.6**	**61.2**	59.6	**59.6**	**63.5**

Natural Images

The performance of our model alongside the baselines on Office-31, Office-Home, and VisDA-C under covariate shift are presented in Table 6.1, 6.2, and 6.4 respectively, where it can be observed that our Penalized Diversity (PD) approach outperforms all UDA and SFDA baselines.

Table 6.4: Target accuracy (%) on VisDA-C under covariate shift.

Method	Source-free	Avg. per-class accuracy
Source-only†	✗	44.6
DAN (Long et al., 2015)	✗	61.1
CDAN (Long et al., 2018)	✗	70.0
MDD (Y. Zhang et al., 2019)	✗	74.6
MCC (Jin et al., 2020)	✗	78.8
Tent (D. Wang et al., 2021)	✓	65.7
SHOT (Liang et al., 2020)	✓	79.6
HDMI (Lao et al., 2021)	✓	82.4
PD	✓	**83.8**

Medical Dataset

The experimental results on the LIDC dataset, given four different domains, are depicted in Table 6.3. The results indicate the effectiveness of PD in comparison with other baselines. Using DBA with WHP increases the performance from 62.3% to 63.5%.

Digit Dataset

The effect of using weighted label entropy in our modified MI maximization objective in the experiments on digit datasets is presented in Table 6.5. Following (Azizzadenesheli et al., 2019), we used two strategies, namely *Tweak-One shift* and *Minority-Class shift* with a probability value of p to create a label distribution shift in each of the datasets. In *Tweak-One shift* (L_t), one of the classes is randomly selected whereas, in *Minority-Class shift* (L_m), a subset of classes (in our experiments, 5 out of 10 classes) is chosen randomly. Then the proportion of the chosen class(es) in the target domain changes by value p, i.e. keeping only p% of the samples in the chosen class (see Fig. 6.3 (b) and (c) for the distribution of labels in each dataset). As demonstrated in Table 6.5, using an estimation of target label distribution as weights in MI objective mitigates the impact of label shift. The improvement is more notable in L_m experiment (more than 8% improvement over OSTAR). In these experiments, covariate shift is also present.

Table 6.5: Target accuracy ($\%$) on digit datasets with no label shift (N_l), tweak-one (L_t), and minority-class (L_m) label distribution shift with $p = 0.1$ (source → target).

Method	Strategy	Source-free	M→U	M→N	U→M	U→N	N→M	N→U	Avg.
OSTAR (Kirchmeyer et al., 2022)†	N_l	×	96.5	33.2	**98.4**	34.0	**98.4**	90.3	75.1
MARS (Rakotomamonjy et al., 2022)†		×	92.3	40.1	96.8	**38.6**	95.4	88.2	75.2
SHOT (Liang et al., 2020)†		✓	88.9	45.9	93.2	29.7	95.4	88.5	73.6
HDMI (Lao et al., 2021)†		✓	95.2	49.6	95.0	26.5	96.2	93.1	75.9
PD		✓	**96.9**	**50.1**	95.6	26.4	96.6	**97.6**	**77.2**
OSTAR (Kirchmeyer et al., 2022)†	L_t	×	94.6	37.9	**98.3**	27.1	**97.4**	85.6	73.5
MARS (Rakotomamonjy et al., 2022)†		×	97.4	44.4	96.2	**44.6**	90.9	89.0	**77.0**
SHOT Liang et al. (2020)†		✓	84.5	46.2	89.3	30.6	89.5	83.6	70.6
HDMI (Lao et al., 2021)†		✓	89.9	48.7	89.6	27.3	90.5	87.2	72.2
PD		✓	**97.6**	**49.1**	92.9	29.1	96.4	**97.0**	**77.0**
OSTAR (Kirchmeyer et al., 2022)†	L_m	×	58.6	37.1	**96.5**	23.1	84.1	67.4	61.1
MARS Rakotomamonjy et al. (2022)†		×	59.9	23.3	92.6	**35.3**	80.0	69.3	60.1
SHOT (Liang et al., 2020)†		✓	57.1	43.9	58.2	28.7	60.9	56.7	50.9
HDMI (Lao et al., 2021)†		✓	62.4	46.0	60.1	25.2	62.6	58.9	52.5
PD		✓	**87.5**	**47.7**	84.3	31.8	**85.0**	**78.6**	**69.2**

Table 6.6: Ablation study on anchor selection under covariate shift: target accuracy ($\%$) on LIDC dataset with DBA under different anchor selection strategies (source → target). 3H represents models with 3 hypotheses. The results are averaged over five different runs.

Method	G→P	G→S	G→T	P→G	P→S	P→T	S→G	S→P	S→T	T→G	T→P	T→S	Avg.
3H-Fixed	68.1	**66.5**	60.7	64.5	64.2	55.5	65.4	**67.8**	58.6	59.7	58.5	59.1	62.4
3H-Random	**69.3**	65.7	**64.5**	64.9	64.3	**55.6**	**66.4**	66.8	57.3	60.4	58.5	**59.9**	62.8
3H-Ensemble	69.2	66.2	59.3	65.3	64.7	54.2	65.9	67.1	56.9	61.0	59.9	**59.9**	62.6
3H-WHP	68.9	65.9	60.9	**65.6**	**65.6**	54.8	65.8	66.9	**66.6**	**61.2**	**59.6**	59.6	**63.5**

6.4.5 Analysis

DBA increases the diversity of source hypotheses

In order to investigate the relative impact on the diversity of introducing separate source hypothesis backbones with the same architecture and using distinct backbone architectures, we follow (Fort, Hu, & Lakshminarayanan, 2019) and measure the source hypotheses' disagreement in function space. More specifically, given a set of target samples X, we compute $\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{M}[f(X;\theta_i) \neq f(X;\theta_j)]$, where N is the total number of target samples, M defines the number of hypotheses, and $f(.)$ indicates the predicted class. Note that in this experiment, we analyze the diversity of the source hypotheses and no adaptation to the target dataset is made. We consider three ways of constructing the ensemble: 1) shared feature extractors, referred to as Shared Backbone (ShB), 2) hypotheses that are given separate feature extractors with the same backbone architecture, referred to as Separate Backbone (SeB), and 3) Our proposed model, DBA, which

Table 6.7: Ablation study on anchor selection under covariate shift: target accuracy (%) on Office-31 dataset with DBA (3 hypotheses) under different anchor selection strategies (source → target).

Method	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
Fixed	94.2	93.8	71.1	98.5	71.0	**99.8**	88.1
Random	94.0	94.1	70.3	98.5	70.4	**99.8**	87.9
Ensemble	95.4	94.2	71.5	98.5	71.1	**99.8**	88.4
WHP	**95.6**	**94.3**	**75.3**	**98.7**	**76.4**	**99.8**	**90.0**

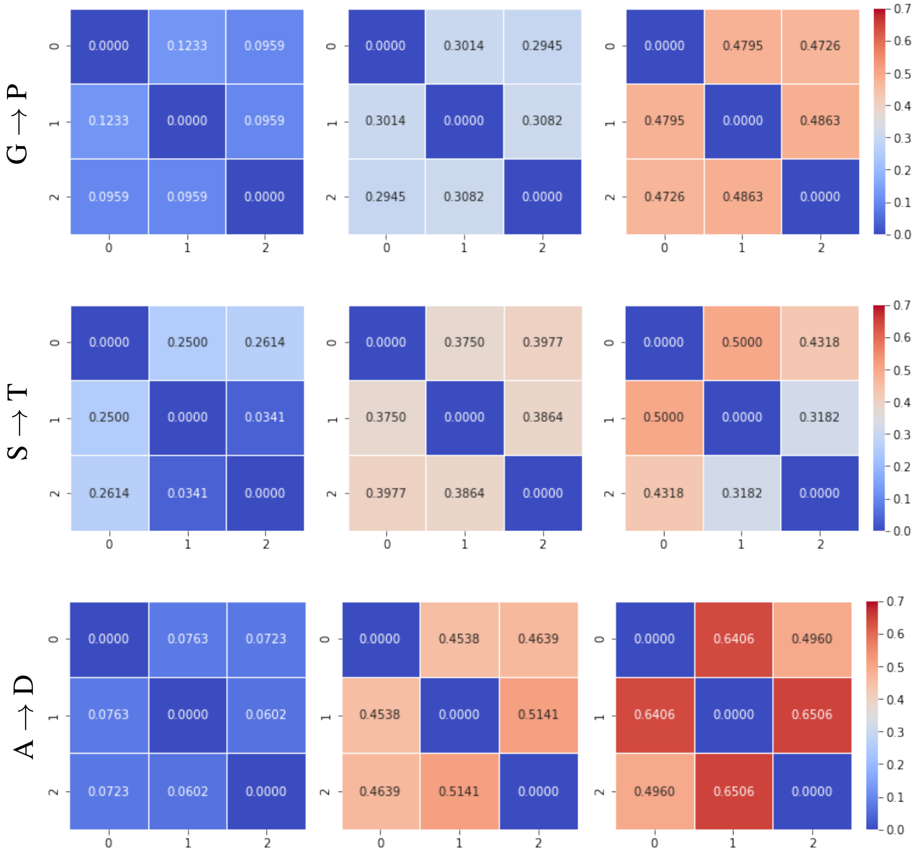
has separate feature extractors. Figure 6.1 shows that simply introducing separate feature extractors for each hypothesis (SeB) leads to a marked increase in diversity compared to sharing a feature extractor (ShB), as in HDMI. However, the largest diversity increase comes from the introduction of DBA.

To further examine the diversity of different ways of constructing the ensemble in the target adaption phase, we show one learning curve example for each model using fixed anchor selection (HDMI objective) in Figure 6.2. As expected, ShB leads to the lowest diversity. Introducing separate backbones (SeB and DBA) induces diversity that leads to an increase in performance. Furthermore, adding a regularizer as the combination of DBA and WHP seems to enable one of the hypotheses to find its way out of a local minimum, underscoring the synergistic impact of Penalized Diversity.

WHP mitigates the negative influence of weak hypotheses

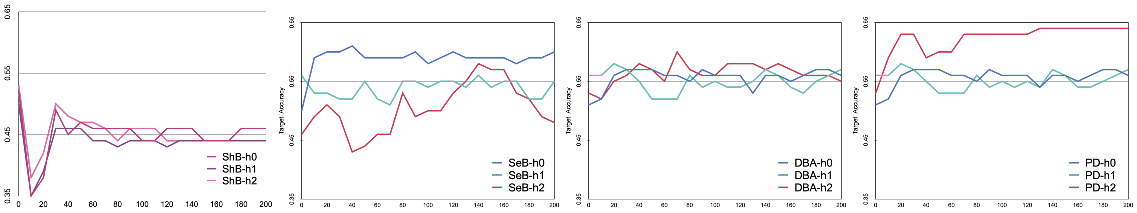
We studied the effect of anchor selection in the target hypothesis disparity regularization. Assuming source and target hypotheses have separate backbones, the performance of our model under fixed, random, ensemble (average), and WHP anchor selection strategies are presented in Table 6.6. It can be observed from Table 6.6 in several experiments, such as S → T, in the presence of weak performing hypothesis, while an ensemble anchor without WHP is subject to convergence towards weak hypotheses, random anchor selection might mitigate this issue partially. However, our results suggest that WHP, through the penalization of outlier hypotheses, provides the most efficient protection against the negative impact of weak hypotheses by assigning them lower weights in the ensemble anchor.

Results on Office-31 dataset with three hypotheses indicate similar findings (see Table 6.7).



(a) ShB (b) SeB (c) DBA

Figure 6.1: **Diversity of the source hypotheses** based on the choice of feature extraction backbone on G → P (first row) and S → T (second row) from LIDC, and A → D from Office-31 datasets using disagreement of predictions between three hypotheses. *Left plot*: three hypotheses with shared features extractors (ShB). *Middle plot*: three hypotheses with separate feature extractors (no weight sharing) (SeB). *Right plot*: three hypotheses with distinct backbone (DBA). All the models were trained with the same random initialization.



(a) ShB (b) SeB (c) DBA (d) DBA + WHP

Figure 6.2: **Diversity of the target hypotheses during adaptation** based on the choice of feature extraction backbone on S → T (LIDC). (a): three hypotheses that share feature extractor (ShB). (b): three hypotheses with separate feature extractors (SeB). (c): DBA with three independent hypotheses. (d): PD = DBA + WHP with three independent hypotheses (DBA).

Given three hypotheses, in all of the fixed, random, and ensemble anchor selection strategies, the impact of a weak hypothesis is inevitable in the overall performance. The results also indicate the instability of random strategy. For instance, in A → W, randomly choosing an anchor improved the performance in comparison with fixed selection, while in D → A, the model seems to converge toward the weak hypothesis. Our results on both natural and medical domains state that using WHP helps to mitigate the effects of weak hypotheses.

Table 6.8: Ablation study on different components of PD = DBA + WHP under covariate shift. Target accuracy (%) on LIDC dataset using three hypotheses (source → target). The results are averaged over five different runs.

Method	G→P	G→S	G→T	P→G	P→S	P→T	S→G	S→P	S→T	T→G	T→P	T→S	Avg.
ShB + Fixed	67.1	66.6	**64.6**	65.4	64.2	54.6	**66.2**	65.1	54.6	60.7	**60.0**	58.7	62.3
SeB + Fixed	66.7	64.4	57.3	64.5	63.3	53.2	65.2	66.4	55.5	60.4	59.9	57.8	61.2
DBA + Fixed	68.1	**66.5**	60.7	64.5	64.2	**55.5**	65.4	**67.8**	58.6	59.7	58.5	59.1	62.4
DBA +WHP	**68.9**	65.9	60.9	**65.6**	**65.6**	54.8	65.8	66.9	**66.6**	**61.2**	59.6	59.6	**63.5**

Penalized Diversity relies on the synergy of DBA and WHP

We ablate different components of the proposed Penalized Diversity (PD) for test-time adaptation performance on the LIDC dataset in Table 6.8. Table 6.8 shows that using a *Fixed* anchor selection, as done in HDMI, can lead to catastrophic failure cases due to error accumulation towards a weak hypothesis, which deems *Fixed* a poor choice for anchor selection. It is important to note that the increase in diversity seen in SeB and DBA (Fig. 6.1) results in worse performance when proper regularization is lacking (SeB + Fixed and DBA + Fixed). It is only when WHP is introduced that we can mitigate the probability of convergence towards weak hypotheses.

Weighted MI mitigates label distribution shift

From Table 6.5, we observe that the performance of SHOT and HDMI dropped by more than 20% in L_m experiment in comparison with the covariate shift only experiment (N_l) indicating the incapability of these models to perform under label distribution shift. Similarly, OSTAR and MARS which both designed to tackle label shift and unlike PD have access to the source data during adaptation, had more than 14% drop in their performance in L_m experiment. While our target estimation obtained from pseudo-labels is prone to errors, it significantly mitigates the catastrophic

Table 6.9: Target accuracy (%) on digit datasets with minority-class (L_m) label distribution shift with $p = 0.1$ (source → target). In these experiments, PD-NWMI refers to PD without the weighted MI maimization, and PD-T refers to PD with the true target class proportions as compared to class proportion generated via pseudo-labels.

Method	M→U	M→N	U→M	U→N	N→M	N→U	Avg.
PD-NWMI	65.9	46.5	61.3	25.3	68.5	62.7	55.0
PD	87.5	47.7	84.3	31.8	85.0	78.6	69.2
PD-T	92.4	52.4	85.1	31.8	87.1	80.7	71.6

impact of label distribution shift by only 8% performance drop. The effect of our modified MI maximization is remarkable in L_t experiment where there is only 0.2% drop in the performance of PD in comparison with no label shift (N_l). It should be noted that our earlier experiments showed that applying W to both label entropy, $H(\hat{Y}^T)$, and conditional entropy, $H(\hat{Y}^T \mid X^T)$, of MI maximization is no better than applying it solely to the label entropy.

Estimated class proportions via pseudo-labels closely represent true class proportions in UDA

To evaluate the effect of weighted MI maximization in the performance of PD under label distribution shift, we compare the performance of PD with and without weighted MI maximization (PD-NWMI) on *Minority-Class shift* experiment. In this experiment, we choose 5 classes out of 10 in the target domain and changed their proportions by $p = 0.1$. Table 6.9 demonstrates a significant improvement ($\%14$) on using estimated target class proportions under label distribution shift.

We further experiment on *Minority-Class shift* with the actual target class proportions. It can be observed from Table 6.9 that the performance of PD using the estimated target class proportions (PD) is close to true target class proportions (PD-T) implying the effectiveness of pseudo-labeling.

6.4.6 Calibration Analysis

It has been shown that diverse ensemble models lead to the best-calibrated uncertainty estimators (Lakshminarayanan et al., 2017). To evaluate the effect of diversity in PD from the calibration perspective, we compute the uncertainty and calibration of PD with Brier score (Brier et al., 1950) and Expected Calibration Error (ECE) (Naeini et al., 2015) and compare it with two other unsupervised SFDA models.

For this experiment, we consider natural and synthetic datasets under covariate and label distribution shifts. It can be observed from Table 6.10, that PD performs well in terms of calibration metrics under covariate shift in natural dataset. We also compare the performance of SFDA models on digit datasets with covariate shift only and with both covariate and label distribution shifts. For the experiment with both covariate and label shifts, we compute the calibration metrics in *Minority-Class Shift* with $p = 0.1$. As seen from Table 6.10, changing the proportion of classes in the target domain not only negatively impacts the transferability of the other two unsupervised SFDA models but also worsens these models' calibration. However, PD with a weighted MI maximization performs significantly better in terms of both performance and calibration after the introduction of label shift.

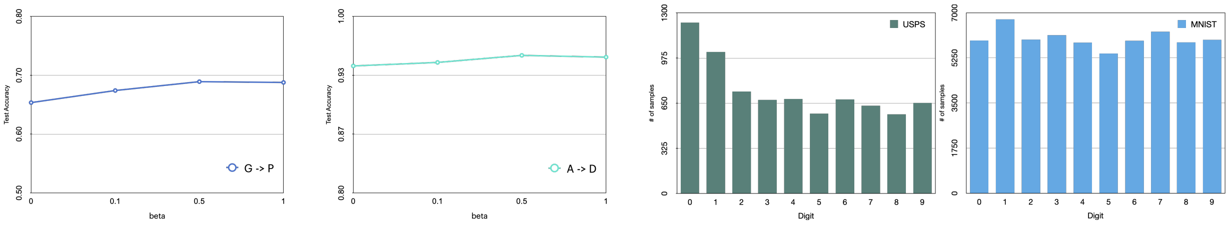
Table 6.10: Calibration estimations for source-free domain adaptation models on the target domains A → D, Office-31, and M → U from digit dataset. Here N_l and L_m represent covariate shift only and covariate shift plus label distribution shift respectively. For the label distribution shift, we consider *Minority-Class shift* with $p = 0.1$. $*$ indicates calibration results reported from the original paper.

Model	Dataset	Shift	Target acc.	Brier Score ↓	ECE ↓
SHOT (Liang et al., 2020)	A → D	N_l	93.1	0.1246	0.0039
HDMI (Lao et al., 2021)*			94.4	0.0961	0.0031
PD			95.6	**0.0771**	**0.0024**
SHOT (Liang et al., 2020)	M → U	N_l	88.9	0.2170	0.0072
HDMI (Lao et al., 2021)			95.2	0.0926	0.0030
PD			96.9	**0.0567**	**0.0011**
SHOT (Liang et al., 2020)	M → U	L_m	57.1	0.8432	0.0279
HDMI (Lao et al., 2021)			62.4	0.7417	0.0246
PD			87.5	**0.2467**	**0.0082**

6.4.7 Sensitivity Analysis

WHP is robust to hyper-parameter selection

To investigate the sensitivity of our model to the hyperparameter β, we conduct a set of experiments on A→D (Office-31) and G→P (LIDC) with three hypotheses and summarize the results in Fig. 6.3. For this experiment, we fix $\alpha = 0.3$ for LIDC and $\alpha = 1$ for Office-31. Setting $\beta = 0$ reduces to solely maximizing mutual information. Fig 6.3 (a) and (b) show that introducing target training with WHP improved the performance in comparison with mutual information maximization ($\beta = 0$). It is seen from the figure that despite the difference between the two domains (natural and medical), increasing the WHP contribution in target training improves the performance.



(a) Model sens. (LIDC)

(b) Model sens. (Office-31)

(c) USPS class dist.

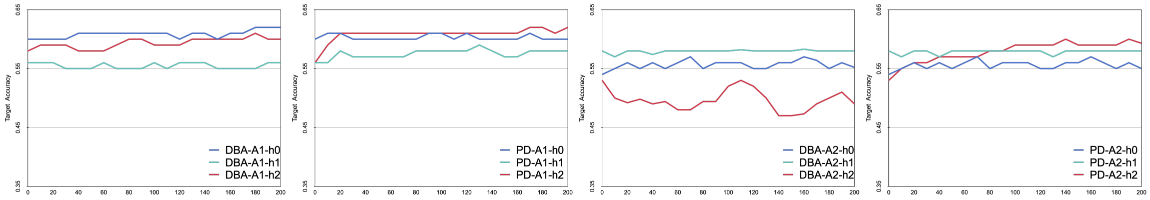
(d) MNIST class dist.

Figure 6.3: The two left plots show β sensitivity in G→P (LIDC) and A→D (Office-31) with three hypotheses. The two right plots indicate class distribution in USPS and MNIST datasets. Note that the class distribution of MNIST-M is the same as MNIST.

6.4.8 Ablation Study

Choice of Different Architectures on PD

We study the effects of different architectural designs on the performance of PD as well as the diversity of the hypotheses. We compare two different choices of architectures for DBA. In this study, we simply consider different depths of a network as different backbones of PD (refers as A1). However, to investigate the performance of PD under totally different architectures, we consider a combination of SqueezeNet (Iandola et al., 2016) and ResNet in PD (refers as A2) with three hypotheses on LIDC. From Figure 6.4, we can observe that three hypotheses with entirely different architectures also improve the diversity. However, DBA without a proper regularizer creates uncontrolled diversity as shown in Figure 6.4(c). The experimental results presented in Table 6.11 show that imposing diversity on the model through entirely different architectural designs (i.e. A2) also leads to improvement in comparison with ShB with similar backbone architectures (from 62.3% to 62.8%).



(a) DBA (A1) (b) PD (A1) (c) DBA (A2) (d) PD (A2)

Figure 6.4: **Diversity of the target hypotheses during adaptation using various architecture choices** on T → G from LIDC dataset. A1 represents the first choice of different backbone architectures including different depths of ResNet {10, 18, 10}. A2 represents the second choice of different backbone architectures including ResNet10, ResNet18, and SqueezeNet1.0. (a): DBA with three hypotheses and fixed anchor on the first architecture (A1). (b): PD with three hypotheses on the first architecture (A1). (c): DBA with three hypotheses and fixed anchor on the second architecture (A2). (d): PD with three hypotheses on the second architecture (A2).

Table 6.11: Ablation study on different architectural choices. Target accuracy (%) on LIDC dataset (source → target) under covariate shift with two different choices of architectures; A1 and A2. The results are averaged over five different runs.

Method	G→P	G→S	G→T	P→G	P→S	P→T	S→G	S→P	S→T	T→G	T→P	T→S	Avg.
PD (A1)	**68.9**	**65.9**	60.9	**65.6**	**65.6**	54.8	**65.8**	**66.9**	**66.6**	**61.2**	**59.6**	**59.6**	**63.5**
PD (A2)	66.2	65.3	**64.3**	63.6	65.2	**59.1**	64.7	**66.9**	61.1	60.6	58.4	58.2	62.8

6.5 Conclusion

This chapter shows the benefits of increasing diversity in unsupervised source-free domain adaptation. We increased diversity by introducing separate feature extractors with Distinct Backbone Architectures (DBA) across hypotheses. With the support of experiments on various domains, we show that diversification must be accompanied by proper Weak Hypothesis mitigation through Penalization (WHP). Our proposed Penalized Diversity (PD) stems from the synergy of DBA and WHP. We further modified MI maximization in the objective of PD to account for the label shift problem. Our experiments on natural, synthetic, and medical benchmarks demonstrate how it improves upon the relevant baselines. As for future work, we would like to investigate other ways to promote diversity in the feature space of SFDA models.

6.6 Supplementary Material

6.6.1 Statistical Analysis on LIDC Covariate Shift

The Lung Image Database Consortium (LIDC) (Armato III et al., 2011) consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans. The images are captured in multiple institutions with imaging devices produced by four different manufacturers. It has been shown that images from different institutions as well as hardware differences in data acquisition devices are susceptible to domain shift (also known as covariate shift) (Guan & Liu, 2021; Karani, Chaitanya, Baumgartner, & Konukoglu, 2018; Stacke, Eilertsen, Unger, & Lundström, 2019). We suggested dividing the LIDC dataset into four sub-datasets based on the imaging device manufacturer. Each of these sub-datasets introduces one domain. Aside from the results presented in Table 6.3 of the paper (Sec. 6.4.4, results on LIDC under covariate shift) obtained by the Source-only model, this section provides a statistical argument for the existence of a covariate shift in our suggested approach.

Assuming that each of the sub-datasets (i.e. domain) is different from the others introducing a comparative observational study or experiment, we assess the difference between proportions or means of each two experiments.

Given the performance of the Source-only model on each domain within five different runs, we computed the confidence interval (CI) between each two population proportions using Eq. 31.

$$CI = \text{P difference} \pm \text{SE for Difference} \tag{31}$$

and

$$\begin{aligned} \text{P difference} &= pr_{d_1} - pr_{d_2} \\ \text{SE for Difference} &= \sqrt{(\text{SE}_{p_1})^2 + (\text{SE}_{p_2})^2} \end{aligned} \tag{32}$$

where SE for proportion p_i defines standard error and it is computed as follows:

$$SE_{p_i} = \sqrt{\frac{pr_{d_i}(1 - pr_{d_i})}{N_{p_i}}}$$

where N_{p_i} indicates the total number of samples in each proportion/domain, and pr_{d_i} is the accuracy of

Table 6.12: The confidence interval (CI) on each domain using the Source-only model on LIDC dataset with covariate shift.

Source	Run	G	P	S	T
G	S1	–	**(-0.09, 2.95)**	(-5.66, -0.20)	(17.63, 21.69)
	S2	–	(-4.96, -1.98)	**(-1.71, 3.85)**	(12.53, 16.59)
	S3	–	(0.67, 3.73)	(-5.53, -0.05)	(18.18, 22.24)
	S4	–	(0.36, 3.40)	(-6.15, -0.73)	(18.10, 22.15)
	S5	–	(3.85, 6.95)	(1.22, 6.90)	(19.10, 23.14)
P	S1	**(-2.10, 0.14)**	–	**(-3.32, 0.74)**	(6.10, 12.62)
	S2	(3.15, 6.15)	–	(6.50, 10.48)	(15.25, 21.73)
	S3	(0.76, 3.82)	–	(1.89, 5.89)	(11.13, 17.63)
	S4	(-3.53, -0.51)	–	**(-1.26, 2.70)**	(12.27, 18.77)
	S5	(3.33, 6.45)	–	(1.53, 5.57)	(10.21, 16.72)
S	S1	(-3.17, -3.13)	**(-0.00, 5.70)**	–	(2.39, 8.85)
	S2	(1.07, 4.11)	(-7.61, -2.27)	–	(5.57, 12.01)
	S3	(1.25, 4.27)	(-6.39, -1.05)	–	(9.05, 15.51)
	S4	(-6.10, -3.05)	(-10.85, -5.45)	–	(5.05, 11.55)
	S5	(-5.42, -2.36)	(-5.72, -0.12)	–	**(-5.33, 0.97)**
T	S1	(18.93, 21.97)	(17.57, 23.33)	(18.44, 22.46)	–
	S2	(18.93, 21.97)	(71.57, 23.33)	(18.44, 22.46)	–
	S3	(22.36, 25.36)	(21.0, 26.72)	(21.87, 25.85)	–
	S4	(17.79, 20.85)	(16.44, 22.20)	(17.31, 21.33)	–
	S5	(17.79, 20.85)	(16.44, 22.20)	(17.31, 21.33)	–

Source-only model trained on the proportion p_i.

If two CIs do not overlap, then it can be said that there is a statistically significant difference between the two populations. In another word, if the confidence interval for the difference does not contain zero, we can confirm the existence of covariate shift between two domains.

The highlighted values in Table 6.12 indicate overlaps between two domains on a particular run. As seen from Table 6.12, the highest domain shift is observed between T and the other domains. These findings are also aligned with the results reported in the paper on the LIDC dataset, where the lowest performance is obtained where T is either source or target domain (see Table 6.3). Since in nearly all the experiments, at least four out of five experiments show no overlaps, we conclude that our suggested approach to creating sub-datasets indeed maintains domain shift.

Chapter 7

Conclusion

Inspired by the importance of model generalizability and non-stationary environments of real-world problems that directly impact the generalizability of predictive models, the ability to detect data distribution drift after model deployment and adapting them to perform under these changes become two popular research topics in the machine and deep learning. In this thesis, we address several problems in medical, natural, and synthetic domains from the perspective of distribution shifts, with an emphasis on detecting and addressing them. In the detection phase, the assumption is that the distribution of the data should remain intact, e.g., the process remains fixed in the production line for a specific product, and hence, any sudden change in the distribution returns a red flag, i.e., either a problem in the production line in a factory or an unknown disease in medical diagnosis. Since many of these changes/shifts are non-deterministic and do not follow a specific pattern, i.e., they are either unspecified or have indefinite varieties, having a training set including a set of labeled examples of them is somehow impractical. Thus, we focus on a system to grasp the underlying data structure with no shift; hence, any sample with a different distribution will be the anomaly, and a shift will be detected. Among existing generative models specialized for understanding data distribution, we use the generative adversarial network (GAN) with a scoring function to measure how far each test/target sample can be from the training distribution to be indicated as an anomaly. Inspired by self-supervision, we further focus on designing a GAN to learn from its internal representations using contrastive learning to mitigate catastrophic forgetting and mode collapse.

On the other hand, in many cases, the distribution of the data might change after deployment leading to a performance drop, e.g., a medical diagnosis tool that is trained on a set of examples dominated by a gender might not perform well if the dominant gender after model deployment becomes the minority. Since these types of changes are not anomalies, the model needs to be adapted. Rather than shifts, over-confident neural networks also decay their generalizability. Motivated by ensemble learning that is proven to improve neural network calibration, we study the effect of an ensemble domain adaptation model. We also tend to increase the diversity among ensemble members since an ensemble model with similar members is no better than a single model. Our domain adaptation model mitigates two significant types of shifts, namely covariate and label distribution shifts, under two major assumptions. First, to maintain data privacy and decrease the storage concern, we consider a source-free domain adaptation (SFDA), and second, no annotated data is accessible in the target domain.

Limitations

Chapter 4 Although the proposed model was able to lower the inference time compared with other anomaly detection models based on GAN, it has difficulty learning training distribution with multiple modes.

Chapter 5 In this work, we proposed a model to tackle the problem of learning multiple modes in training distribution from previous work. However, introducing contrastive learning to GANs slightly increases both the training and inference time of the model.

Chapter 6 One of the limitations of this work is that the WHP regularizer performs better when the number of ensemble members is at least three. Besides, diversity in an ensemble model with different architectures needs a proper search in the space of architectures.

Future Perspectives

Reasoning the Shift The problem of explaining the decisions of an anomaly detector, in general, out-of-distribution (OOD) detector, remains largely unexplored. Most current studies focus on improving their performance on their OOD detectors while exploring the explainability, i.e., the reason behind the learner's decision to identify a sample as anomaly/OOD remains neglected. Therefore, one direction for future research is to go beyond shift detection and investigate a human-understandable interpretation method for anomaly detectors.

Diversity in SFDA In this thesis, the diversity in the ensemble has been brought by different architectures of neural networks. Even though this trivial approach is effective, it poses a few questions. How to search for possible architectures? What are the criteria for selecting a particular architecture? Therefore, as future work, we would like to explore other ways, such as contrastive learning to increasing diversities in an ensemble.

Code Availability

All the conducted experiments and the code related to our first research project can be found here. The codes of AD-CGAN and all the related experiments, including the baselines, are available here.

Chapter 8

Appendix

In this chapter, we review the details of each dataset used for experiments on anomaly detection and domain adaptation in Chapter 4, 5, and 6.

8.1 Dataset

In this thesis, several natural, synthetic, and medical dataset has been used. Table 8.1 presents the training and test sizes of all the natural datasets for the task of anomaly detection. We also present the sizes of all the datasets for the domain adaptation task in Table 8.2.

Table 8.1: Training and test sizes of all the natural datasets for our anomaly detection experiments.

Samples	MNIST	FashionMNIST	CIFAR10	CatsVsDogs
Train	60,000	60,000	50,000	20,000
Test	10,000	10,000	10,000	5,000

The images of the USPS dataset are grayscale images of $(16, 16, 1)$. MNIST and FashionMNIST (fMNIST) are grayscale images of $(28, 28, 1)$. MNIST_M is a colored MNIST of images of the same size. CIFAR10 icludes RGB images of $(32, 32, 3)$. CatsVsDogs has RGB images of different sizes where we scaled all the images to $(64, 64, 3)$.

VisDA-C has 12 classes with 152,397 Synthetic images in the source domain and 55,388 Real images in the target domain.

For our anomaly detection experiments, we also considered a medical imaging dataset. Acute Lymphoblastic Leukemia (ALL) dataset (Labati et al., 2011) has 260 images with the size $(257, 257, 3)$. Each

Table 8.2: Training and test sizes of all the natural and synthetic datasets for our domain adaptation experiments.

Dataset	MNIST	MNIST_M	USPS	Office-31	Office-Home	VisDa-C
Total sizes	70,000	70,000	9,298	4,652	15,500	207,785

normal and anomalous class has 130 images.

We considered the Lung Image Database Consortium (LIDC) (Armato III et al., 2011) as a medical dataset in our domain adaptation experiments. We divided the LIDC dataset into four domains based on the manufacturer of the data-capturing device: GE_medical (G), Philips (P), SIEMENS (S), and TOSHIBA (T) with 1976, 346, 838, and 266 samples, respectively. Each of these domains has two classes, healthy and unhealthy.

References

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). Detrac: Transfer learning of class decomposed medical images in convolutional neural networks. *IEEE Access*, *8*, 74901–74913.

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, *2*(1), 1–18.

Antipov, G., Baccouche, M., & Dugelay, J.-L. (2017). Face aging with conditional generative adversarial networks. In *2017 ieee international conference on image processing (icip)* (pp. 2089–2093).

Antorán, J., Allingham, J., & Hernández-Lobato, J. M. (2020). Depth uncertainty in neural networks. *Advances in neural information processing systems*, *33*, 10620–10634.

Argyriou, A., Evgeniou, T., & Pontil, M. (2006). Multi-task feature learning. *Advances in neural information processing systems*, *19*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017, 06–11 Aug). Wasserstein generative adversarial networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 214–223). International Convention Centre, Sydney, Australia: PMLR.

Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... others (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, *38*(2), 915–931.

Ashukha, A., Lyzhov, A., Molchanov, D., & Vetrov, D. (2020). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=BJxI5gHKDr

Ayerdi, B., Savio, A., & Graña, M. (2013). Meta-ensembles of classifiers for alzheimer's disease detection using independent roi features. In *International work-conference on the interplay between natural and artificial computation* (pp. 122–130).

Aytekin, C., Ni, X., Cricri, F., & Aksu, E. (2018). Clustering and unsupervised anomaly detection with l_2 normalized deep auto-encoder representations. In *Proc. of ijcnn* (pp. 1–6).

Azizzadenesheli, K., Liu, A., Yang, F., & Anandkumar, A. (2019). Regularized learning for domain adaptation under label shifts. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=rJl0r3R9KX

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4945–4949).

Ballard, D. H. (1987). Modular learning in neural networks. In *Aaai* (Vol. 647, pp. 279–284).

Bengio, E., Jain, M., Korablyov, M., Precup, D., & Bengio, Y. (2021). Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, *34*, 27381–27394.

Bergman, L., & Hoshen, Y. (2020). Classification-based anomaly detection for general data. In *International conference on learning representations.*

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on machine learning* (pp. 81–88).

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Berlin, Heidelberg: Springer-Verlag.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 acm sigmod international conference on management of data* (pp. 93–104).

Brier, G. W., et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly* *weather review*, *78*(1), 1–3.

Brown, G. (2004). *Diversity in neural network ensembles* (Unpublished doctoral dissertation). University of Birmingham, United Kingdom. (Winner, British Computer Society Distinguished Dissertation Award)

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4960–4964).

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*, *14*, 15.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 1–58.

Chau, M., & Chen, H. (2008). A machine learning approach to web page filtering using content and structure analysis. *Decision Support Systems*, *44*(2), 482–494.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today*, *23*(6), 1241–1250.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, *58*, 101539.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

Chen, T., Zhai, X., Ritter, M., Lucic, M., & Houlsby, N. (2019). Self-supervised gans via auxiliary rotation loss. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12154–12163).

Cheng, Z., Zhu, E., Wang, S., Zhang, P., & Li, W. (2021). Unsupervised outlier detection via transformation invariant autoencoder. *IEEE Access*, *9*, 43991–44002.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 8789–8797).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 1, pp. 886–893).

Dasarathy, B. V., & Sheela, B. V. (1979). A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, *67*(5), 708–713.

Dash, A., Gamboa, J. C. B., Ahmed, S., Liwicki, M., & Afzal, M. Z. (2017). TAC-GAN-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*.

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, *77*(379), 605–610.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., & Kloft, M. (2018). *Anomaly detection with generative adversarial networks.* Retrieved from https://openreview.net/forum?id=S1EfylZ0Z

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09.*

Deng, Z., Luo, Y., & Zhu, J. (2019). Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9944–9953).

Donahue, J., Krähenbühl, P., & Darrell, T. (2017). Adversarial feature learning. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings.* OpenReview.net.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Dou, Q., Ouyang, C., Chen, C., Chen, H., & Heng, P.-A. (2018). Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*.

Du, S. S., Koushik, J., Singh, A., & Póczos, B. (2017). Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, *30*.

Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., & Courville, A. C.

(2017). Adversarially learned inference. In *Proc. of iclr.*

Elson, J., Douceur, J. R., Howell, J., & Saul, J. (2007). Asirra: a captcha that exploits interest-aligned manual image categorization. *CCS*, *7*, 366–374.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition*, *58*, 121–134.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115–118.

Fawcett, T., & Provost, F. (1997). Adaptive fraud detection. *Data mining and knowledge discovery*, *1*(3), 291–316.

Fernandes, K., & Cardoso, J. S. (2019). Hypothesis transfer learning based on structural model similarity. *Neural Computing and Applications*, *31*(8), 3417–3430.

Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148–156).

Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.).* USA: Academic Press Professional, Inc.

Gadelha, M., Maji, S., & Wang, R. (2017). 3d shape induction from 2d views of multiple objects. In *2017 international conference on 3d vision (3dv)* (pp. 402–411).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Ganapathiraju, A., Hamaker, J., & Picone, J. (2000). Hybrid svm/hmm architectures for speech recognition. In *Sixth international conference on spoken language processing.*

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180–1189).

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., … Lempitsky, V.

(2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096–2030.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 580–587).

Golan, I., & El-Yaniv, R. (2018). Deep anomaly detection using geometric transformations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., & Schölkopf, B. (2016). Domain adaptation with conditional transferable components. In *International conference on machine learning* (pp. 2839–2848).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (http://www.deeplearningbook.org)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672–2680).

Guan, H., & Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning* (pp. 1321–1330).

Guo, X., Yao, Q., Kwok, J., Tu, W., Chen, Y., Dai, W., & Yang, Q. (2020). Privacy-preserving stacking with application to cross-organizational diabetes prediction. In *Federated learning* (pp. 269–283). Springer.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 ieee computer society conference on computer vision and pattern recognition (cvpr'06)* (Vol. 2, pp. 1735–1742).

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., … Nakayama, H. (2018). GAN-based synthetic brain MR image generation. In *Proc. of isbi* (pp. 734–738).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904–1916.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *International conference on learning representations.*

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278–282).

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, *16*(5), 550–554.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360.*

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. of cvpr* (pp. 1125–1134).

Jiang, J., & Zhai, C. (2007). Instance weighting for domain adaptation in nlp..

Jianliang, M., Haikun, S., & Ling, B. (2009). The application on intrusion detection based on k-means cluster algorithm. In *2009 international forum on information technology and applications* (Vol. 1, pp. 150–152).

Jin, Y., Wang, X., Long, M., & Wang, J. (2020). Minimum class confusion for versatile domain adaptation. In *European conference on computer vision* (pp. 464–480).

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 133–142).

Jolicoeur-Martineau, A. (2019). The relativistic discriminator: a key element missing from standard GAN. In *International conference on learning representations.*

Karani, N., Chaitanya, K., Baumgartner, C., & Konukoglu, E. (2018). A lifelong learning approach to brain MR segmentation across scanners and protocols. In *International conference on medical image computing and computer-assisted intervention* (pp. 476–484).

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 4401–4410).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 8110–8119).

Keller, T. A., & Welling, M. (2021). Topographic vaes learn equivariant capsules. *Advances in Neural Information Processing Systems*, *34*, 28585–28597.

Kemker, R., McClure, M., Abitino, A., Hayes, T., & Kanan, C. (2018). Measuring catastrophic forgetting in neural networks. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Kingma, D. P., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, *28*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirchmeyer, M., Rakotomamonjy, A., Bezenac, E. d., & gallinari, p. (2022). Mapping conditional distributions for domain adaptation under generalized target shift. In *International conference* *on learning representations.* Retrieved from https://openreview.net/forum?id=sPfB2PI87BZ

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning* (pp. 3519–3529).

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images* (Unpublished master's thesis). Computer Science Department, University of Toronto.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuzborskij, I., & Orabona, F. (2013). Stability and hypothesis transfer learning. In *International conference on machine learning* (pp. 942–950).

Kuzborskij, I., & Orabona, F. (2017). Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, *106*(2), 171–195.

Labati, R. D., Piuri, V., & Scotti, F. (2011). All-idb: The acute lymphoblastic leukemia image database for image processing. *2011 18th IEEE International Conference on Image Processing*, 2045-2048.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

Lao, Q., Havaei, M., Pesaranghader, A., Dutil, F., Jorio, L. D., & Fevens, T. (2019). Dual adversarial inference for text-to-image synthesis. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 7567–7576).

Lao, Q., Jiang, X., & Havaei, M. (2021). Hypothesis disparity regularized mutual information maximization. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 8243–8251).

Latecki, L. J., Lazarevic, A., & Pokrajac, D. (2007). Outlier detection with kernel density functions. In *International workshop on machine learning and data mining in pattern recognition* (pp. 61–75).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision* (pp. 319–345). Springer.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., … others (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4681–4690).

Lee, K. S., Tran, N.-T., & Cheung, N.-M. (2021). Infomax-gan: Improved adversarial image generation via information maximization and contrastive learning. In *Proceedings of the ieee/cvf winter conference on applications of computer vision* (pp. 3942–3952).

Lee, Y., Yao, H., & Finn, C. (2022). Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*.

Leung, K., & Leckie, C. (2005). Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the 28h australasian conf. on computer science-volume 38* (pp. 333–342).

Li, Y., Murias, M., Major, S., Dawson, G., & Carlson, D. (2019). On target shift in adversarial domain adaptation. In *The 22nd international conference on artificial intelligence and statistics* (pp. 616–625).

Li, Y., Wang, N., Shi, J., Liu, J., & Hou, X. (2016). Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*.

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning* (pp. 6028–6039).

Lipton, Z., Wang, Y.-X., & Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning* (pp. 3122–3130).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60–88.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international* *conference on data mining* (pp. 413–422).

Liu, X., Xing, F., Yang, C., El Fakhri, G., & Woo, J. (2021). Adapting off-the-shelf source segmenter for target medical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 549–559).

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural networks*, *12*(10), 1399–1404.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning* (pp. 97–105).

Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2018). Conditional adversarial domain adaptation. *Advances in neural information processing systems*, *31*.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations.*

Louzada, F., & Ara, A. (2012). Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. *Expert Systems with Applications*, *39*(14), 11583–11592.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh ieee international conference on computer vision* (Vol. 2, pp. 1150–1157).

Lucic, M., Kurach, K., Michalski, M., Gelly, S., & Bousquet, O. (2018). Are GANs created equal? a large-scale study. In *Proc. of nips* (pp. 700–709).

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*.

Mahadevan, V., Li, W., Bhalodia, V., & Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *Proc. of cvpr* (pp. 1975–1981).

Mahapatra, D., Bozorgtabar, B., & Garnavi, R. (2019). Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, *71*, 30–39.

Mahapatra, D., & Ge, Z. (2020). Training data independent image registration using generative adversarial networks and domain adaptation. *Pattern Recognition*, *100*, 107109.

Mathieu, M., Couprie, C., & LeCun, Y. (2016). Deep multi-scale video prediction beyond mean square error. In *International conference on learning representations.*

Minsky, M., & Papert, S. (1969). 1969), perceptrons. *Cambridge, MA: MIT Press*, *18*, 19.

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

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*(37), 870–877.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations.*

Mohan, A., Chen, Z., & Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. In *Proceedings of the learning to rank challenge* (pp. 77–89).

Mu, X., Lu, J., Watta, P., & Hassoun, M. H. (2009). Weighted voting-based ensemble classifiers with application to human face recognition and voice recognition. In *2009 international joint conference on neural networks* (pp. 2168–2171).

Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Twenty-ninth aaai conference on artificial intelligence.*

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. (2011). Reading digits in natural images with unsupervised feature learning. *NIPS*.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems* (pp. 841–848).

Nguyen, T., Raghu, M., & Kornblith, S. (2021). Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=KJNcAkY8tY4

Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., & Shen, D. (2017). Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 417–425).

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

Ouyang, C., Kamnitsas, K., Biffi, C., Duan, J., & Rueckert, D. (2019). Data efficient unsupervised domain adaptation for cross-modality image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 669–677).

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, *32*.

Pagliardini, M., Jaggi, M., Fleuret, F., & Karimireddy, S. P. (2022). Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Park, Y., & Kellis, M. (2015). Deep learning for regulatory genomics. *Nature biotechnology*, *33*(8), 825.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065–1076.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2536–2544).

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559–572.

Peng, X., Usman, B., Kaushik, N., Wang, D., Hoffman, J., & Saenko, K. (2018). Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 2021–2026).

Quang, D., & Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, *44*(11), e107–e107.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset shift* *in machine learning*. Mit Press.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. of iclr.*

Rafiee, L., & Fevens, T. (2020). Unsupervised anomaly detection with a gan augmented autoencoder. In *International conference on artificial neural networks* (pp. 479–490).

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, *32*.

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on machine learning* (pp. 759–766).

Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M. E., Berar, M., & Courty, N. (2022). Optimal transport for conditional domain matching and label shift. *Machine Learning*, *111*(5), 1651–1670.

Rame, A., & Cord, M. (2021). DICE: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *International conference on learning representations.*

Redko, I., Courty, N., Flamary, R., & Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In *The 22nd international conference on artificial intelligence and statistics* (pp. 849–858).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. In *Proceedings of the 33rd international conference on international conference on machine learning - volume 48* (p. 1060–1069). JMLR.org.

Reed, S. E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., & Lee, H. (2016). Learning what and where to draw. In *Proc. of nips* (pp. 217–225).

Rifai, S., Vincent, P., Muller, X., Glorot, X., & Bengio, Y. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on international conference on machine learning* (p. 833–840). Madison, WI, USA:

Omnipress.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial intelligence review*, *33*(1), 1–39.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., ... Kloft, M. (2018). Deep one-class classification. In *International conference on machine learning* (pp. 4393–4402).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning internal representations by error propagation. In *Neurocomputing: Foundations of research* (p. 673–695). Cambridge, MA, USA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *European conference on computer vision* (pp. 213–226).

Salakhutdinov, R., & Hinton, G. (2009, 16–18 Apr). Deep boltzmann machines. In D. van Dyk & M. Welling (Eds.), *Proceedings of the twelth international conference on artificial intelligence and statistics* (Vol. 5, pp. 448–455). Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, *5*(2), 197–227.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proc. of ipmi* (pp. 146–157).

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., Platt, J. C., et al. (1999). Support vector method for novelty detection. In *Nips* (Vol. 12, pp. 582–588).

Shui, C., Li, Z., Li, J., Gagné, C., Ling, C. X., & Wang, B. (2021). Aggregating from multiple target-shifted sources. In *International conference on machine learning* (pp. 9638–9648).

Sill, J., Takács, G., Mackey, L., & Lin, D. (2009). Feature-weighted linear stacking. *arXiv preprint arXiv:0911.0460*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Sohn, K., Li, C.-L., Yoon, J., Jin, M., & Pfister, T. (2021). Learning and evaluating representations for deep one-class classification. In *International conference on learning representations.*

Sood, R., Topiwala, B., Choutagunta, K., Sood, R., & Rusu, M. (2018). An application of generative adversarial networks for super resolution medical imaging. In *2018 17th ieee international conference on machine learning and applications (icmla)* (pp. 326–331).

Stacke, K., Eilertsen, G., Unger, J., & Lundström, C. (2019). A closer look at domain shift for deep learning in histopathology. *arXiv preprint arXiv:1909.11575*.

Stickland, A. C., & Murray, I. (2020). Diverse ensembles improve calibration. *arXiv preprint arXiv:2007.04206*.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., & Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, *20*.

Swati, Z. N. K., Zhao, Q., Kabir, M., Ali, F., Ali, Z., Ahmed, S., & Lu, J. (2019). Brain tumor classification for MR images using transfer learning and fine-tuning. *Computerized Medical Imaging and Graphics*, *75*, 34–46.

Tachet des Combes, R., Zhao, H., Wang, Y.-X., & Gordon, G. J. (2020). Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, *33*, 19276–19289.

Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10781–10790).

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*(1), 45–66.

Tran, N.-T., Tran, V.-H., Nguyen, B.-N., Yang, L., & Cheung, N.-M. M. (2019). Self-supervised gan: Analysis and improvement with multi-class minimax game. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Turner, K., & Oza, N. C. (1999). Decimated input ensembles for improved generalization. In *Ijcnn'99. international joint conference on neural networks. proceedings (cat. no. 99ch36339)* (Vol. 5, pp. 3069–3074).

Vanetti, M., Binaghi, E., Carminati, B., Carullo, M., & Ferrari, E. (2010). Content-based filtering in on-line social networks. In *International workshop on privacy and security issues in data mining and machine learning* (pp. 127–140).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Venkateswara, H., Eusebio, J., Chakraborty, S., & Panchanathan, S. (2017). Deep hashing network for unsupervised domain adaptation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5018–5027).

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning* (pp. 1096–1103).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 ieee computer society conference on computer vision and pattern recognition. cvpr 2001* (Vol. 1, pp. I–I).

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *Advances in neural information processing systems* (pp. 613–621).

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=uXl3bZLkr3c

Wang, F., Han, Z., Gong, Y., & Yin, Y. (2022). Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7151–7160).

Wang, X., & Schneider, J. G. (2015). Generalization bounds for transfer learning under model shift. In *Uai* (pp. 922–931).

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.

Wolterink, J. M., Leiner, T., Viergever, M. A., & Išgum, I. (2017). Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, *36*(12), 2536–2545.

Woodland, P. C., & Povey, D. (2002). Large scale discriminative training of hidden markov models for speech recognition. *Computer Speech & Language*, *16*(1), 25–47.

Wu, J., Zhang, C., Xue, T., Freeman, B., & Tenenbaum, J. (2016). Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems* (pp. 82–90).

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xiong, L., Póczos, B., & Schneider, J. G. (2011). Group anomaly detection using flexible genre models. In *Proc. of nips* (pp. 1071–1079).

Xu, R., Li, G., Yang, J., & Lin, L. (2019). Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1426–1435).

Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017). Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning* (pp. 3861–3870).

Yang, C., Wu, Z., Zhou, B., & Lin, S. (2021). Instance localization for self-supervised detection pretraining. In *Cvpr*.

Yang, S., van de Weijer, J., Herranz, L., Jui, S., et al. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, *34*, 29393–29405.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., & Do, M. N. (2017).

Semantic image inpainting with deep generative models. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5485–5493).

Yu, Y., Min, X., Zhao, S., Mei, J., Wang, F., Li, D., … Li, S. (2020). Dynamic knowledge distillation for black-box hypothesis transfer learning. *arXiv preprint arXiv:2007.12355*.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on machine learning* (p. 114).

Zaidi, S., Zela, A., Elsken, T., Holmes, C., Hutter, F., & Teh, Y. W. (2020). Neural ensemble search for performant and calibrated predictions. *arXiv preprint arXiv:2006.08573*, *2*, 3.

Zaidi, S., Zela, A., Elsken, T., Holmes, C. C., Hutter, F., & Teh, Y. (2021). Neural ensemble search for uncertainty estimation and dataset shift. *Advances in Neural Information Processing Systems*, *34*.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., & Chandrasekhar, V. R. (2018). Efficient GAN-based anomaly detection. *arXiv preprint arXiv:1802.06222*.

Zenati, H., Romain, M., Foo, C.-S., Lecouat, B., & Chandrasekhar, V. (2018). Adversarially learned anomaly detection. In *2018 ieee international conference on data mining (icdm)* (pp. 727–736).

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. of iccv* (pp. 5907–5915).

Zhang, K., Schölkopf, B., Muandet, K., & Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International conference on machine learning* (pp. 819–827).

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666).

Zhang, Y., Liu, T., Long, M., & Jordan, M. (2019). Bridging theory and algorithm for domain adaptation. In *International conference on machine learning* (pp. 7404–7413).

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5810–5818).

Zhao, Q., Zhang, Y., Friedman, D., & Tan, F. (2015). E-commerce recommendation with personalized promotion. In *Proceedings of the 9th acm conference on recommender systems* (pp. 219–226).

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 665–674).

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of iccv* (pp. 2223–2232).

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations.*