Methodologies to Improve Fidelity and Reliability of Deep Learning Models for Real-World Deployment

by

Vivek Sivaraman Narayanaswamy

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2023 by the
Graduate Supervisory Committee:

Andreas Spanias, Chair
Jayaraman J. Thiagarajan
Visar Berisha
Cihan Tepedelenlioglu

ARIZONA STATE UNIVERSITY

May 2023

ABSTRACT

The past decade witnessed the success of deep learning models in various applications of computer vision and natural language processing. This success can be predominantly attributed to the (i) availability of large amounts of training data; (ii) access of domain aware knowledge; (iii) *i.i.d* assumption between the train and target distributions and (iv) belief on existing metrics as reliable indicators of performance. When any of these assumptions are violated, the models exhibit brittleness producing adversely varied behavior.

This dissertation focuses on methods for accurate model design and characterization that enhance process reliability when certain assumptions are not met. With the need to safely adopt artificial intelligence tools in practice, it is vital to build reliable failure detectors that indicate regimes where the model must not be invoked. To that end, an error predictor trained with a self-calibration objective is developed to estimate loss consistent with the underlying model. The properties of the error predictor are described and their utility in supporting introspection via feature importances and counterfactual explanations is elucidated.

While such an approach can signal data regime changes, it is critical to calibrate models using regimes of inlier (training) and outlier data to prevent under- and over-generalization in models *i.e.,* incorrectly identifying inliers as outliers and vice-versa. By identifying the space for specifying inliers and outliers, an anomaly detector that can effectively flag data of varying semantic complexities in medical imaging is next developed.

Uncertainty quantification in deep learning models involves identifying sources of failure and characterizing model confidence to enable actionability. A training strategy is developed that allows the accurate estimation of model uncertainties and

its benefits are demonstrated for active learning and generalization gap prediction. This helps identify insufficiently sampled regimes and representation insufficiency in models.

In addition, the task of deep inversion under data scarce scenarios is considered, which in practice requires a prior to control the optimization. By identifying limitations in existing work, data priors powered by generative models and deep model priors are designed for audio restoration. With relevant empirical studies on a variety of benchmarks, the need for such design strategies is demonstrated.

DEDICATION

To

My parents, teachers and the almighty!

ACKNOWLEDGMENTS

I would like to express the first and foremost regard to my doctoral advisor Dr. Andreas Spanias for his guidance, support and trust. Believing in me and my capabilities, he continuously encouraged me to pursue all my research ambitions. Forming my academic collaborations that allowed me to explore a broad spectrum of research would not have been possible without his support.

I would also like to extend my deepest gratitude to my mentor Dr. Jayaraman J. Thiagarajan (LLNL) for his unconditional support, encouragement, time and guidance, and for always rekindling the fire in me to achieve and learn more. Under his tutelage, I got the chance to work on diverse research topics and focus on real-world applications. Moreover, I sincerely thank him for providing me the opportunity to intern with LLNL for several summers helping me build my skills. He has been a person to whom I look up to and seek inspiration from, in time of need.

I sincerely thank my committee members Dr. Visar Berisha and Dr. Cihan Tepedelenlioglu for their useful insights, guidance and feedback on my research. I would like to thank the SenSIP Center, ASU for supporting me through multiple NSF grants (1646542, 2019068) on solar array monitoring and industry collaborations. Special thanks to Dr. Raja Ayyanar and Dr. Devarajan Srinivasan (Poundra LLC) for their useful discussions and insights on my projects. I would acknowledge Silvia Garre (NXP), Lalo Canales (NXP) and Gowtham Muniraju (NXP) for encouraging me to be creative about new ideas and methods that can be used at the intersection of AI and sensor design.

I am extremely grateful to Toni Mengert and Robina Sayed who have been my go-to people for my academic and administrative needs. I would like to thank all my colleagues Rakshith Subramanyam, Kowshik Thopalli, Sunil Rao, Sameeksha

Katoch, Uday Shanthamallu, Mohit Malu, Kristen Jaskie, Huan Song, Deep Pujara, Bhavikkumar Patel, Araditha Sharma, Glen Uehera, Rajesh Sathya Kumar and Max Yarter at ASU for their help and support. In addition, I would also like to thank Rushil Anirudh (LLNL), Deepta Rajan (Microsoft) and Yamen Mubarka (LLNL) for their useful feedback and guidance on several research ventures.

Most importantly, my PhD could not have transpired without the emotional support and backing from my friends and family. Their love and support have energized me to achieve my goals. In particular, I would like to thank my parents Uma and Narayanaswamy, my sister Indira, my brother in-law Balaji and my wonderful niece Aarabhi for their constant motivation. I would also like to specially thank my best friends Balaji Eevunni, Srivatsan, Srinath, Niyantha, Tej and Sreeram for always being on my side and guiding me through the right path in life. Finally, my eternal gratitude to the almighty for making all this happen!

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

In recent years, deep neural networks (DNNs) (Shanthamallu and Spanias 2022; Shanthamallu et al. 2017; V. Narayanaswamy et al. 2023; Rao et al. 2020) and artificial intelligence (AI) tools have become the *modus-operandi* in a variety of fields including computer vision (Goodfellow et al. 2014; Karras, Laine, and Aila 2019; Karras et al. 2020; Karras et al. 2021), speech processing (Baevski et al. 2020; Rao et al. 2021) and natural language processing (Vaswani et al. 2017; Devlin et al. 2019). As a result, these tools have been rapidly adopted in applications such as healthcare (Hosny et al. 2018; Young et al. 2020), speech retrieval systems (Amodei et al. 2016) and autonomous driving (Rao and Frtunikj 2018). The widespread adoption of deep neural networks (DNNs) has resulted in their predictions being trusted and used to make important decisions in fields such as medical diagnosis and prognosis (Hosny et al. 2018) and criminal justice (Shah, Bhagat, and Shah 2021). Moreover, there has been increased emphasis from policy makers and tech-investors (AI Weekly 2022) to build sophisticated AI models and tools (Karras et al. 2021; Zhuang Liu et al. 2022) with a focus on improving performance on known benchmarks and open-source data.

Deep neural networks have seen tremendous success due to a number of assumptions that are built into their training and deployment processes. One assumption is the availability of large amounts of training data to learn useful heuristics to generalize well to unseen data. Typical examples include large-scale language models such as BERT (Devlin et al. 2019) and GPT-3 (Brown et al. 2020). Another assumption is the availability of domain-specific knowledge to guide training and improve model

quality. Examples include task specific loss functions (Thiagarajan, Venkatesh, and Rajan 2020), priors and inductive biases (Ulyanov, Vedaldi, and Lempitsky 2018), or knowledge encodings (Subramanyam et al. 2022; Narayanaswamy, Thiagarajan, et al. 2019). The third factor is that the DNNs are usually evaluated under *closed-world* conditions (Fei and Liu 2016) where the target data is considered to be independent and identically distributed with respect to the training data. In other words, there exists no apparent distribution shifts between the train and target data which is seldom the case in practice. Finally, the evaluation of DNNs often relies on pre-existing metrics (Hendrycks and Gimpel 2017) that are believed to be good indicators of performance.

1.1 Motivation

Under conditions where the assumptions underlying the use of DNNs are not fully met, even state-of-the-art DNNs have been shown (Guo et al. 2017) to be brittle and to exhibit poor reliability, generalization, and process fidelity. This means that they may not perform well when applied to tasks or data that differ significantly from those they were trained on, and may produce unreliable or unexpected results. One common method for evaluating the reliability and robustness of DNNs is to estimate confidence under distribution shifts where the target distribution is different from that of the training distribution. For example, if a DNN is trained on a dataset of skin lesion images from a hospital and then applied to a dataset of skin lesions procured from another hospital, there may be a distribution shift due to the variability in patient demographics. Estimating confidence under such shifts is critical to identify whether the model may be safely adopted and predictions are trustworthy. It can also shed

light on the sensitivities of a model to anomalies or out-of-distribution (OOD) data. However, DNN predictions have been shown to be poorly calibrated (Guo et al. 2017) under distribution shifts. This means that the predicted confidence does not match the true data likelihoods, and hence not meeting the expectation of the end-user. Moreover, DNNs also produce high confidence scores for anomalous data instead of preferably abstaining from predictions. Confidence in deep models is conventionally measured using the *softmax* scores obtained from the final layer of classification models. While this metric is useful to provide an estimate of model accuracy, it is based on the inherent stochasticities of model training and hence may not accurately reflect true confidences. This motivates the requirement for: (i) designing predictive models that are well calibrated and sensitive to outliers; (ii) developing a reliable metric to characterize model confidence and (iii) building anomaly detectors capable of accurately identifying a variety of distribution shifts.

While there exists a multitude of problems where existing DNNs offer limited flexibility, inverse problems (Anirudh et al. 2020; Narayanaswamy, Subramanyam, et al. 2022; Narayanaswamy, Katoch, et al. 2019) offer unique challenges. The goal of inverse problems is to recover the input given the corrupted or transformed observations. Examples include image restoration, where the goal is to recover a clear image from a degraded or noisy version, and source separation, where the goal is to recover individual sources from a mixture of signals. These types of problems are often challenging because the transformation or corruption applied to the input data can be complex and difficult to invert. Moreover, deep inverse problems are further complicated by the requirement of high-fidelity recovery under data-scarce scenarios. Consequently, typical solutions often rely on the use of *priors* or regularizers (Ulyanov, Vedaldi, and Lempitsky 2018; Shah and Hegde 2018) that constrain the solution

space of the recovered inputs to prevent trivial convergence and handle the inherently ill-posed nature of these problems. It has been shown that existing approaches for deep inversion lead to poor source recovery as they rely on weak, statistical priors (Vivek Narayanaswamy et al. 2020) on the structure of the inferred solutions. As a result, we require the careful design of advanced priors and training strategies to achieve improved recovery and scalability.

1.2 Problem Statement

This dissertation focuses on exploring real-world problems where it is imperative to appropriately design models, training strategies and reliable metrics to improve overall model fidelity, reliability and deployability.

1.2.1 Designing Reliable Failure Detectors for Deep Deterministic Models

A vital step towards safely deploying predictive models in practice is to ensure that (i) the models produce accurate estimates of true confidences; (ii) they do not generalize to data regimes where the training data provide no meaningful evidence and (iii) they are well calibrated to detect distribution shifts. Existing approaches rely on off-the-shelf metrics (Guo et al. 2017) or explicit uncertainty estimators (Gal and Ghahramani 2016) that are inherently hard to calibrate. This emphasizes the need to design reliable surrogates for uncertainty estimation that are inherently well-calibrated, easy to incorporate into any deep model, can handle distribution shifts, sensitive to data variations and importantly indicate model failure. Moreover, for increasing trust in the model under *open-world* conditions, it is important to understand the strengths and

weaknesses through model introspection. Conventionally, model introspection can be performed via feature importance explanations or counterfactual generation. Currently available explanation methods are limited by computational complexity (Lundberg and Lee 2017), difficulty in adapting to distribution shifts, and reliance on adversarial training data or generative models (Verma, Dickerson, and Hines 2020). To address these limitations, it is crucial to leverage well-calibrated failure detectors that are consistent with the model and can accurately provide feature relevance scores and assist in generating plausible counterfactual explanations.

1.2.2 Calibrating Classifiers for Improving Anomaly Detection

Accurately detecting out-of-distribution (OOD) data with varying levels of semantic and covariate shifts with respect to the in-distribution (ID) data is critical for deployment of safe and reliable models. Although one can resort to training a reliable failure indicator to identify anomalies, it is important to well-calibrate the OOD detector with data from inlier and outlier distributions in order to control under- and over-generalization in deep models and improve anomaly detection. Conventionally, pixel space inlier augmentations and curated outlier datasets are used for calibration. However, existing strategies either adversely affect anomaly detection or require representative datasets which may not be available in applications such as medical imaging. To that end, it is important to identify the appropriate inlier and outlier specification required for calibration together with the space chosen for their implementation.

1.2.3 Uncertainty Quantification in Deep Models for Improved Actionability

While failure indicators from deterministic models are reliable, they do not identify the causes of model failure (measurement noise, model errors, label noise, optimization errors) and lack actionability. To that end, uncertainty quantification (UQ), a popular statistical tool can be adopted in deep models to better support the rigorous characterization of confidence/failure. UQ allows the estimation of aleatoric or data uncertainty and epistemic or model uncertainties that arise in the training pipeline. Existing approaches for estimating such uncertainties rely on methods that are computationally complex such as Bayesian Neural Networks (Lampinen and Vehtari 2001) and deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) or those that produce poorly calibrated uncertainty estimates (Gal and Ghahramani 2016). This motivates the need to design and develop scalable uncertainty estimators that not only reliably characterize confidence but also identify regimes of improper sampling for active learning (Yoo and Kweon 2019). Moreover, as an additional safety measure while adopting deep models in practice, it is crucial to estimate generalization accuracy of a pre-trained model under distribution shifts. While UQ can aid in characterizing shifts between two data domains, it is important to extend such principles to uncover model representation uncertainties and produce richer metrics indicative of performance.

1.2.4 Designing Data and Model Priors for Deep Inverse Problems

Restoration tasks such as denoising and source separation are inverse problems that require the high-fidelity estimation of the inputs given the corrupted/mixture observations. Under limited data settings, training an exclusive DNN (Luo and Mesgarani

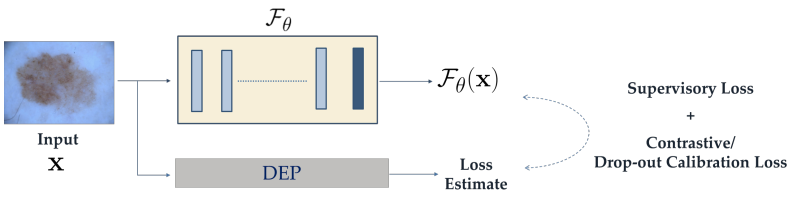
2019) to robustly and reliably solve inversion is significantly challenging. In such scenarios, deep inversion relies on the careful specification of priors that place plausible assumptions on the nature of the inferred solutions. However, existing work relies on off-the-shelf prior designs that place weak priors on the inferred solution space (Bishop and Nasrabadi 2006). Examples include (i) Independent Component analysis that promotes the non-gaussianity of separated sources and (ii) Total variation (Ulyanov, Vedaldi, and Lempitsky 2018) which only preserves local smoothness. To that end, it becomes critical to design task and modality specific priors, and loss functions that can effectively guide the inverse optimization.

1.3 Contributions

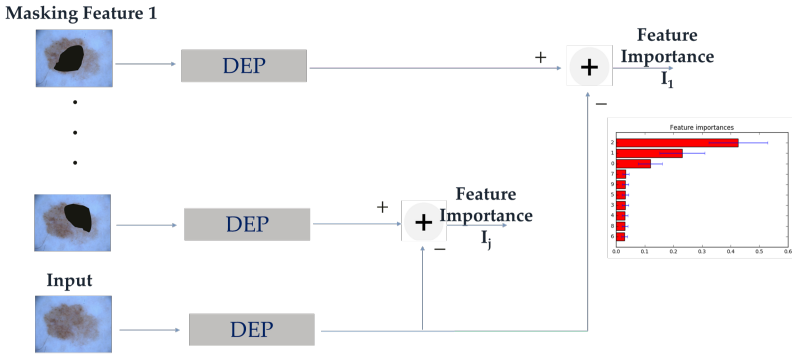
This section outlines the contributions to each problem statement and the organization of the dissertation. Figures 1, 2, 3, and 4 visually depict each of the contributions.

1.3.1 Direct Error Predictors as Reliable Failure Indicators

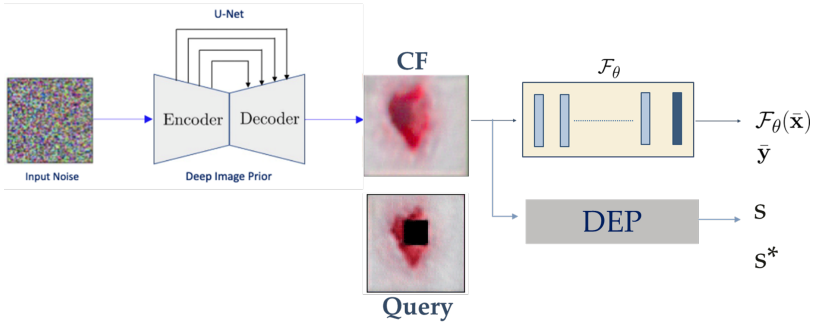
In chapter 3 (Vivek Narayanaswamy et al. 2021; Narayanaswamy, Rajan, et al. 2022), a direct error predictor (DEP) to produce reliable model failure indicators is proposed. The importance of such a design amidst existing work is discussed. The strategies for training DEPs alongside the predictive model based on the contrastive or dropout calibration based objectives are shown. The methods for directly estimating the error or failure indicator is described. The improvement in generalization behavior of the underlying model trained jointly with DEP is showcased on a challenging



(a) Training Direct Error Predictors (DEPs)



(b) Feature Importance Estimation with DEP



(c) Counterfactual Synthesis with DEP

Figure 1. Training Methodology and Utility of Direct Error Predictors (DEPs). (a) The DEP Is Trained Alongside the Predictor Model Driven by a Contrastive/Drop-out Calibration Based Objective. Such a Strategy Preserves the Ranking of the Loss Estimates with Respect to the Underlying Model Ensuring Consistency and Fidelity. (b) Using the Pre-trained DEP, We Can Obtain the Feature Importance Scores for a given Image. Here, the Input Features Are Masked Sequentially and the Relevance Scores Can Be Calculated as Shown. (c) For a given Query Image, the DEP Can Effectively Guide an Inverse Optimization Such as Counterfactual Synthesis in Order to Produce Plausible Images That Belong to the Original Training Data Manifold.

skin lesion benchmark (Codella et al. 2018). In addition, the enhancement in outlier detection of the model jointly trained with the DEP is demonstrated across a variety of

semantic and modality data distribution shifts. In addition to producing well-calibrated uncertainties, the other properties of DEP namely: (i) image corruption detection; (ii) prototypical sample identification; (iii) sensitivities to pixel-level/feature-level variations are elucidated with examples.

1.3.2 Feature Importance Estimation using Direct Error Predictors

In chapter 4 (JJ. Thiagarajan et al. 2021), the utility of DEPs is extended to support feature importance estimation. The method PRoFILE (Producing Robust Feature Importances using Loss Estimates) which is a novel feature importance estimation method that addresses different bottlenecks of existing approaches is proposed. The training paradigm of DEPs is first revisited. The Granger Causality objective (Granger 1969) is adopted for estimating feature relevance scores based on the feature masking style explanation methods using the pre-trained DEP. The impact of PRoFILE in efficiently capturing relevant data features is demonstrated over a wide variety of synthetic and real-world benchmarks of both image and non-image data. Finally, the robustness of PRoFILE in preserving the feature importance scores under distribution shifts is demonstrated. Moreover, the importance of the proposed approach to produce high-fidelity scores without any sophisticated training strategies unlike (Lakkaraju, Arsov, and Bastani 2020) or computational complexities is discussed.

1.3.3 Counterfactual Synthesis using Deep Model Inversion

The properties of the DEP to identify image corruptions, offer pixel-level sensitivities and identifying prototypes are utilized in chapter 5 for synthesizing counterfactual

images. The focus of chapter 5 (V. Narayanaswamy, Thiagarajan, and Spanias 2021; Jayaraman Thiagarajan et al. 2021) is on the case where only the pre-trained classifier is available and not the actual training data to aid counterfactual generation. While the problem of using deep models to synthesize images from the training distribution has been explored (Mordvintsev, Olah, and Tyka 2017), the goal is to develop a deep inversion approach for generating counterfactual explanations for a given query image. Despite their effectiveness in conditional image synthesis, it is extensively shown that existing deep inversion methods are insufficient for producing meaningful counterfactuals. A method called DISC (Deep Inversion for Synthesizing Counterfactuals) is proposed which improves upon deep inversion by using stronger image priors, incorporating a novel manifold consistency objective enforced using DEPs, and adopting a progressive optimization strategy. The different components in the synthesis pipeline are elaborated upon with their pros and cons. It is found that, in addition to producing visually meaningful explanations, the counterfactuals from DISC are effective at learning classifier decision boundaries and are robust to unknown test-time corruptions. Empirical studies across different natural and medical image benchmarks are conducted to validate the proposed approach.

1.3.4 Inlier and Outlier Specifications for Calibrating OOD Detectors

In chapter 6 (Narayanaswamy et al. 2022; Narayanaswamy, Mubarka, et al. 2022), the problem of anomaly detection in medical imaging is presented. While, there are extensive inference time scoring functions (Liu et al. 2020) to accurately identify anomalies, the importance of calibrating the detector with regimes of inlier and outlier data for such an application is elucidated. The calibration objective that allows the

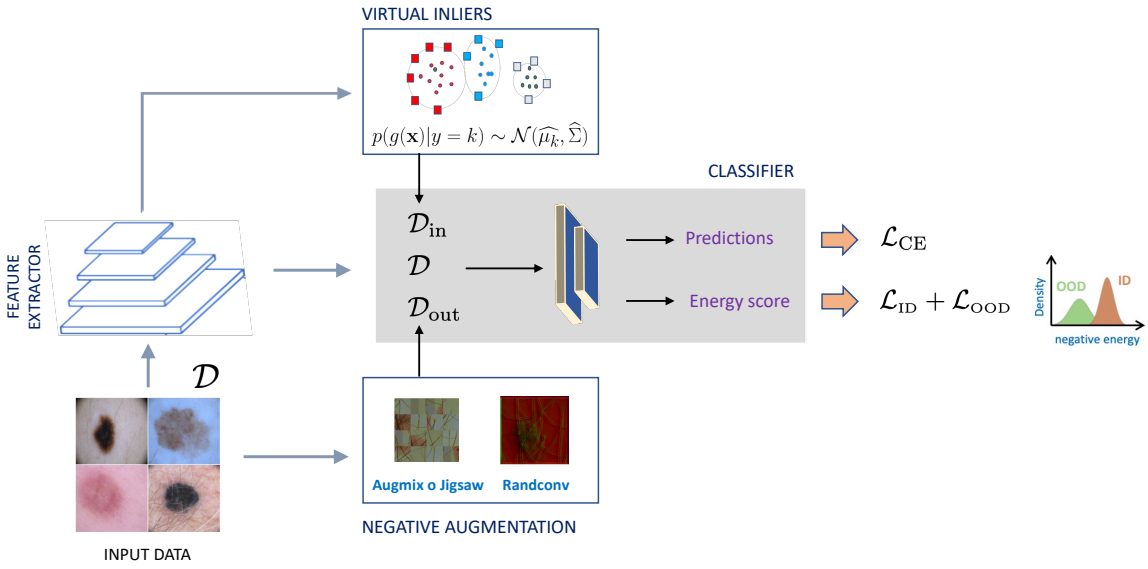
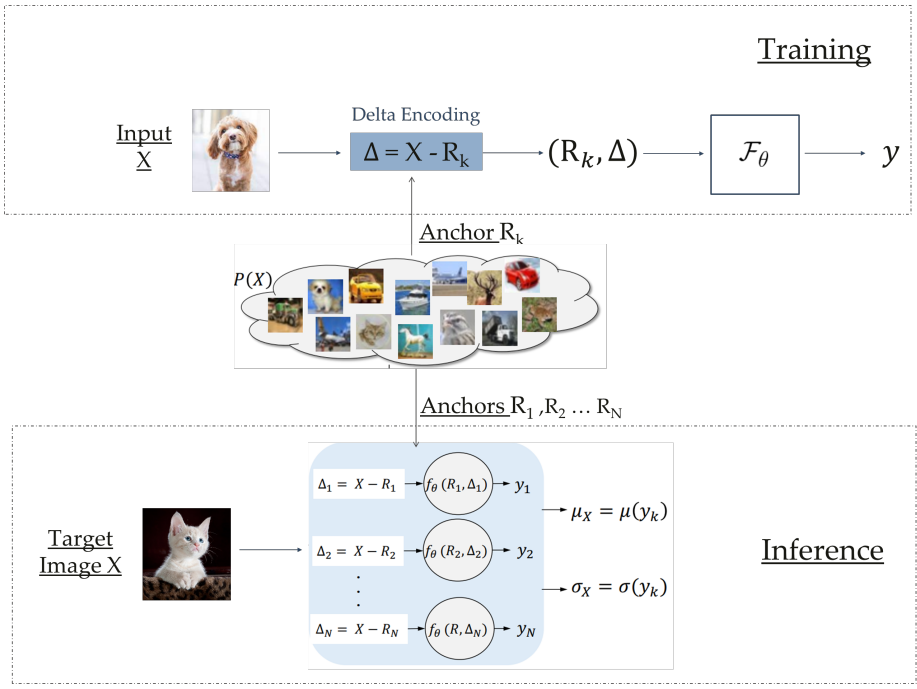
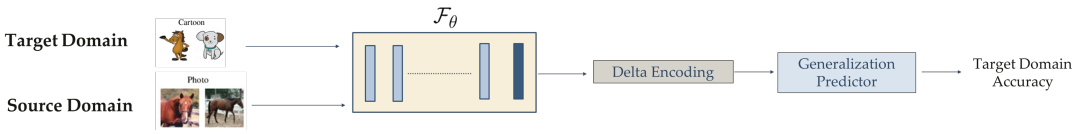


Figure 2. Calibrating Classifiers for Improving OOD Detection Through the Specification of Synthetic Inliers and Outliers. We Focus on Energy-based OOD Detectors for Deep Models and Explore the Design of Synthetic Augmentations. We Propose to Utilize Inliers Synthesized in the Latent Space and Outliers Synthesized in the Pixel Space to Calibrate the Classifier and Improve OOD Detection. Our Training Pipeline Consistently Leads to High-fidelity Detectors in Both near and Far OOD Settings, Without Compromising the Test Accuracy in Comparison with Existing Baselines.

detector to not compromise on the inlier accuracy while also rejecting examples from OOD regimes is presented. The conventional strategies for realizing this dual objective namely inlier specification via pixel space augmentations and outlier specification using a curated set of OOD data are discussed along with their shortcomings. The performance of the existing methods on medical OOD detection is shown to highlight the requirement of better augmentation specifications. To that end, it is found that inlier specification through augmentations constructed in the latent space together with exposure to diverse, synthetic pixel space outliers obtained from the training data are essential for medical OOD detection. Through a rigorous empirical study on medical imaging benchmarks, significant performance gains (15% to 35% in AUROC) over existing approaches are reported under both semantic and modality shifts.



(a) Training and Inference with Δ-UQ



(b) Using Δ-UQ for Predicting Generalization Gaps in Deep Models

Figure 3. Overview of Δ-UQ and Its Utility in Uncovering Representation Uncertainties in Deep Models. (a) During Training, Every Input Sample Is Combined with a Randomly Selected Anchor from the Training Distribution to Form an Encoding Which Is Then Used to Train the Classifier. In Every Iteration, the Same Input Sample Gets Associated with Different Anchors and Consistency Is Enforced in Prediction Irrespective of the Anchor. During Inference, the Mean and Uncertainty of the Prediction Is Obtained by Marginalizing the Impact of Different Anchors. (b) On a Pre-trained Model, the Principles of Δ-UQ Are Applied to Train a Post-hoc Accuracy Estimator That Predicts the Generalization on a Target Dataset.

1.3.5 Single Model Epistemic Uncertainty Estimation via Stochastic Data Centering

In chapter 7 (J. J. Thiagarajan et al. 2022; Anirudh 2021; Anirudh and Thiagarajan 2022), uncertainty quantification for deep models is introduced for the rigorous

characterization of model confidence and identifying individual sources of error. The chapter explicitly focuses on estimating *epistemic* uncertainties or model uncertainties from neural networks. The bottlenecks of using existing approaches are discussed and the need for a scalable uncertainty estimation model is established. It is found that an ensemble of neural networks with the same weight initialization, trained on datasets that are shifted by a constant bias, gives rise to slightly inconsistent trained models, where the differences in predictions are a strong indicator of epistemic uncertainties. Using the neural tangent kernel (NTK) (Jacot, Gabriel, and Hongler 2018) framework, it is demonstrated that this phenomena occurs in part because the NTK is not shift-invariant. Since this is achieved via a trivial input transformation, the chapter shows that this behavior can be approximated by training a single neural network using a technique called Δ-UQ that estimates uncertainty around prediction by marginalizing out the effect of the biases during inference. The chapter extensively covers the litmus tests adopted for testing episetemic uncertainties namely outlier rejection, calibration under distribution shift, and sequential design optimization of black box functions. It is also demonstrated that Δ-UQ's uncertainty estimates are superior to many current methods on a variety of benchmarks.

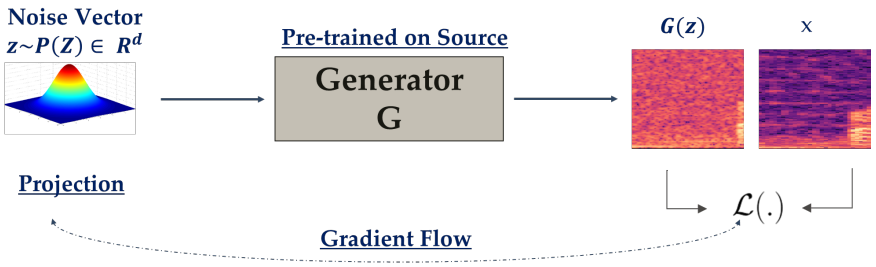
1.3.6 Uncovering Representation Uncertainties for Generalization Gap Prediction in Deep Models

In chapter 8 (Narayanaswamy, Anirudh, et al. 2022), a novel strategy is proposed for directly predicting accuracy on unseen target data by leveraging the principle of Δ-UQ to uncover representation uncertainty in predictive models. This uncertainty refers to the inability of the underlying model to accurately represent (in the feature

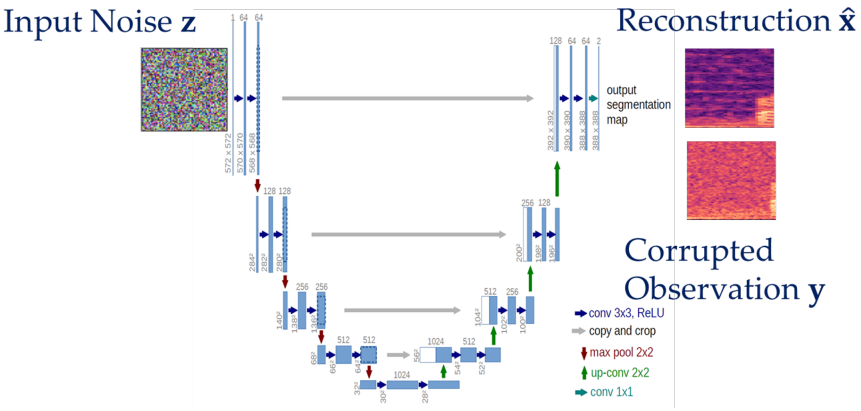
space) distribution shifts of the original data manifold. The effectiveness of Δ-UQ in characterizing and detecting domain shifts is exploited for predicting the generalization gap. A training pipeline is proposed where the Δ-UQ based encoding is used to train the generalization gap predictor on a suitable calibration dataset. Inferencing with the pre-trained generalization gap predictor is also elaborately provided. Experiments on datasets with a wide variety of real-world distribution shifts are considered and it is shown that the proposed approach outperforms existing baselines (Guillory et al. 2021).

1.3.7 Generative Priors for Audio Source Separation

Chapter 9 (Vivek Narayanaswamy et al. 2020) focuses on the problem of unsupervised audio source separation which requires the specification of an appropriate prior to control the inversion. The drawbacks of the existing priors are elucidated and the need for next generation data priors is presented. Next-gen data priors that are powered by generative models are described and the projected gradient descent (PGD) style inferencing with generative priors is illustrated. The PGD based inversion with a single generative prior is extended to a multiple generative prior based inversion. In particular, the effectiveness of source-specific generative priors for recovering the constituents of an unlabeled mixture is demonstrated. The process of choosing an appropriate prior and the need for carefully designed spectral domain loss functions to obtain high fidelity source estimates is also described. It is shown that the proposed approach offers superior flexibility and can efficiently handle a varying number of known sources in a given mixture, in contrast to standard supervised approaches which require re-training or extensive fine-tuning.



(a) Inversion with Generative Priors



(b) Inversion with Deep Model Priors

Figure 4. Deep Inversion with Data Priors Powered by Generative Models or Structural Priors Powered by Carefully Tailored DNN Architectures. (a) For Inversion with Generative Priors, the Generator Is Sampled Using the Latent Code $\mathbf{z}$ to Synthesize an Audio Spectrogram. The Synthesis Is Compared with the Ground Truth Spectrum Using a Suitable Objective Function. The Latent Code $\mathbf{z}$ is Updated Using Projected Gradient Descent in an Effort to Estimate the Code That Best Matches the given Observation. (b) For a given Observation, Deep Model Priors Optimize the Parameters of an Untrained DNN Whose Structure Provides a Prior to the Space of Audio.

1.3.8 Designing Deep Model Priors for Audio Restoration

The ill-posed task of unsupervised audio restoration is considered in chapter 10 (V. S. Narayanaswamy, Thiagarajan, and Spanias 2021). In lieu of generative priors, this chapter focuses on the design of model priors that rely on the careful choice of an untrained, tailored neural network to control inversion. The existing gaps in current model prior designs are identified and it is found that such approaches are

either computationally intensive or require sophisticated engineering of convolutional kernels. A high fidelity, yet efficient model prior design is proposed to solve such inverse problems. In particular, it is identified that including adaptive dilated convolutions and dense connections to extract useful multi scale features significantly improves the fidelity as well as the computational complexity for ill-posed restoration tasks such as denoising, in-painting, and source separation across different audio benchmarks. This design improves the robustness of the model to sampling rate changes and enables complex temporal modeling.

Chapter 2

BACKGROUND

In this chapter, we provide an overview of the foundational concepts that are relevant to the problems addressed in this dissertation. We begin by examining the importance of accurately characterizing confidence in deep learning models, including model calibration (Guo et al. 2017) and the pros and cons of different metrics for representing confidence. We then introduce the problem of out-of-distribution detection in the context of multi-class classification, describing the current strategies and scoring functions used, as well as the various types of out-of-distribution data that a detector must be able to handle. We then turn to the importance of uncertainty quantification (J.J Thiagarajan et al. 2020) for accurately characterizing confidence and making informed decisions. We describe the various sources of uncertainty and the current methods for estimating them. Finally, we explore the topic of inverse problems (Vivek Narayanaswamy et al. 2020) and the existing approaches for addressing them.

2.1 Characterizing Confidence in Predictive Models

The effectiveness of modern predictive models based on deep neural networks (DNNs) is often evaluated using metrics such as accuracy, which measures the correctness of predictions. These metrics use prediction likelihood distributions, such as *softmax* probabilities, to determine the predicted labels. While accuracy can be a useful way to evaluate model performance, it does not provide any information about why

the model made a particular prediction or whether the model has relied on shortcuts to make its predictions. Additionally, when DNNs are deployed in real-world scenarios where the distribution of data may differ from the training distribution (Anirudh 2021), their predictions are often poorly calibrated (Guo et al. 2017). In other words, the *softmax* estimates do not necessarily reflect the model's confidence and can produce overly confident probability estimates for data not seen during training (Snoek, Larochelle, and Adams 2012; Hendrycks and Gimpel 2017). These measures of predictive likelihoods are dependent on the model's training dynamics and may not reflect the underlying data likelihoods or model stochasticities. More importantly, they may not align with the end-user's understanding of the underlying process. This motivates the need (i) to effectively represent model confidence and hence provide useful model failure indicators and, (ii) to train well-calibrated predictive models.

2.1.1 Confidence Calibration

A model is said to be well-calibrated when the model confidence or the predictive probability of correctness aligns with the model accuracy or the observed probability of correctness (Tran et al. 2022).

Consider, the supervised multi-class classification setting in DNNs. Let the input $\mathbf{x} \in \mathcal{X}$ and label $y \in \mathcal{Y} = \{1, \ldots, K\}$ be random variables following a joint distribution $P(\mathbf{x}, y) = P(y|\mathbf{x})\ P(\mathbf{x})$. Let f be a neural network with $f(\mathbf{x}) = (\hat{y}, \hat{P})$, where $\hat{y}$ is the class prediction and $\hat{P}$ is its associated confidence also referred to as the probability of correctness.

Ideally, we expect the confidence estimate $\hat{P}$ to be calibrated, which means that $\hat{P}$ represents the true probability of correctness. For instance, if a model is well-calibrated,

a prediction made with a probability of 0.8 should be correct approximately 80% of the time. This is important because if a model is poorly calibrated, its predicted probabilities may not accurately reflect the true likelihood of the predicted classes. This can lead to incorrect decision-making or overconfidence in the model's predictions.

Mathematically, we can define *calibration* as

$$\mathbb{P}\left(\hat{y} = y|\hat{P} = p\right) = p, \quad \forall p \in [0, 1] \tag{2.1}$$

where the probability is over the joint distribution. However, achieving perfect calibration is not possible due to a variety of reasons such as finite data availability and model dynamics. On this regard, there have been several approximations of measuring calibration in predictive models. For instance, *expected calibration error* (ECE) (Anirudh 2021) measures the difference between the model accuracy and predictive confidence. ECE essentially divides the model predictions on the data samples into n-equally spaced bins and estimates the true accuracy of the samples within the bins. A lower value of ECE signifies that the model is well-calibrated and vice-versa. Other methods include estimating the negative log-likelihood (Lakshminarayanan, Pritzel, and Blundell 2017) which is the same as the cross-entropy loss used for training DNNs.

2.1.2 Commonly Adopted Metrics for Characterizing Confidence

There exists a number of metrics for quantifying model confidence which were originally designed to reveal the relationship between the underlying data distributions and the prediction likelihoods. For instance, *maximum softmax probability* which is

defined as

$$\hat{P}_{\mathbf{x}}^{\max} = \max_{y} \quad \frac{\texttt{exp}(z_{\mathbf{x},y})}{\sum_{y'=1}^{K} \texttt{exp}(z_{\mathbf{x},y'})}, \tag{2.2}$$

is the most common choice to represent confidence. Here, $z_{\mathbf{x},y}$ represents the y^{th} logit (unscaled log probability) from the output layer for sample $\mathbf{x}$.

Another popularly adopted metric for characterizing model confidence is *predictive entropy*. It is given by,

$$H[P(y|\mathbf{x})] = -\sum_{y \in \mathcal{Y}} P(y|\mathbf{x}) \log P(y|\mathbf{x}). \tag{2.3}$$

A higher value of predictive entropy implies that the information about a given sample is evenly distributed among the classes indicating poor model confidence. Other methods of measuring confidence include using the *energy* of a sample. Energy is an unconstrained function that maps every sample $\mathbf{x}$ in the input space to a deterministic scalar. Such functions have been demonstrated to directly relate to the data distribution $p(\mathbf{x})$ (Liu et al. 2020). In modern DNNs, the energy is given by the negative logit g corresponding to a given class label (Liu et al. 2020). Mathematically, it is given by

$$E(\mathbf{x}, y) = -g(\mathbf{x}, y). \tag{2.4}$$

While these measures offer simplicity and ease of adoption, they have been demonstrated to be poorly calibrated under distribution shifts not meeting the expectation of the end-user or evidence provided by the data.

2.2 Out-of-Distribution Detection in Predictive Models

2.2.1 Preliminaries

Out-of-distribution (OOD) detection (Hendrycks and Gimpel 2017; Anirudh and Thiagarajan 2022; Narayanaswamy et al. 2022) is the task of identifying when an ML model or a DNN is making predictions on input data that is significantly different from the training data or in-distribution (ID). This is important because a model that is trained on one distribution of data may not generalize well to other, significantly different distributions, and may produce unreliable or even harmful predictions. OOD detection can be especially important in safety-critical applications, such as medical diagnosis (Hosny et al. 2018) or self-driving cars, where incorrect predictions can have serious consequences. In general, the goal of OOD detection is to identify data that is significantly different from the training distribution, while minimizing the number of false negatives (i.e., incorrectly classifying in-distribution data as OOD) and false positives (i.e., incorrectly classifying OOD data as in-distribution).

2.2.2 Definition

To address this challenging problem, it is necessary to effectively characterize the manifold of ID data, such that the difference between OOD data and the inferred manifold can be used to identify when the model lacks knowledge about out-of-distribution (OOD) data. This is typically achieved by learning a scoring function, $\mathcal{S} : X \to \mathrm{R}$, that can accurately score both ID and OOD samples. Mathematically,

the OOD detector $\mathcal{G}_{OOD}$ can be formulated as,

$$\mathcal{G}_{OOD}(\mathbf{x}; \mathcal{F}_\theta, \tau) = \begin{cases} \text{outlier}, & \text{if } \mathcal{S}(\mathbf{x}, \mathcal{F}_\theta) \leq \tau, \\ \text{inlier}, & \text{if } \mathcal{S}(\mathbf{x}, \mathcal{F}_\theta) > \tau. \end{cases} \tag{2.5}$$

Here, $\mathcal{F}_\theta$ corresponds to the ML model pre-trained on the ID data and τ is a user-defined threshold for detection.

2.2.3 Classes of Out-of-Distribution Data

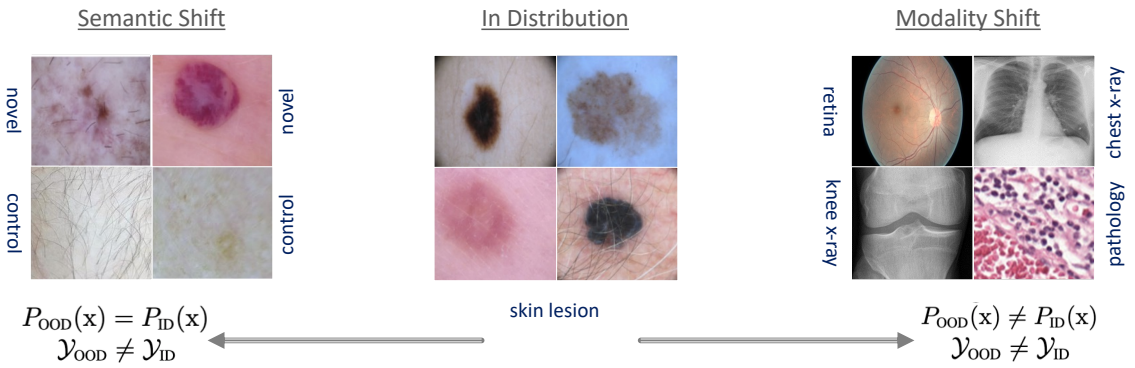


Figure 5. An Illustrative Example from Medical Imaging, the Classes of OOD Data. For an OOD Detector Trained on Skin Lesions, Examples of Images from Novel Disease States or Control Groups Constitute Near OOD Data (Semantic Shifts). On the Other Hand, Examples from Disparate Domains Such as X-rays Constitute Far OOD Data (Modality Shifts).

Out-of-distribution (OOD) data can be broadly classified into two categories based on the similarity of its semantic concepts to the in-distribution (ID) data that a model was trained on: Near OOD and Far OOD. Near OOD data is similar to the ID data, but is different enough to be considered OOD. For example, in medical imaging (Figure 5), Near OOD data may represent a new disease state or class unseen during training. In contrast, Far OOD data is significantly different from the ID data and may come from a completely different distribution or domain. For instance, chest X-ray

images could be considered Far OOD for a model trained on images of skin lesions. It is important to distinguish between Near and Far OOD data because different models and OOD detection methods may be more or less effective at detecting each type. Some strategies may be more sensitive to Near OOD data, while others may be better at detecting Far OOD data. The effectiveness of a given method or model at detecting Near or Far OOD data may depend on various factors such as the complexity of the ID data distribution, the amount of training data available, and the specific characteristics of the Near or Far OOD data.

2.2.4 Existing Methods

There are several methods and scoring functions developed for detecting OOD data at inference time. These include the *softmax* scores (Hendrycks and Gimpel 2017), predictive entropy (Andreas et al. 2021), and energy derived from classification models. Other methods include post-hoc calibration strategies such as Platt scaling (Guo et al. 2017), which use tempered softmax probabilities to score between ID and OOD data based on a validation dataset. More recent methods such as *ODIN* (Liang, Li, and Srikant 2018) and Mahalanobis distance-based scoring functions (K. Lee et al. 2018) use adversarial perturbations on ID data and extensive hyper-parameter tuning on a validation set to temper their respective scoring functions.

While there exists other methods based on using novel scoring functions (Sastry and Oore 2020; Ren et al. 2021) or explicit model uncertainty estimators (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017), they may not be effective in detecting all classes of OOD data namely Near and Far OOD data regimes, which is crucial in applications such as healthcare. One solution to this problem is to

explicitly calibrate the classifier with regimes of carefully curated outlier data using a Outlier Exposure (OE) (Hendrycks, Mazeika, and Dietterich 2018; Thulasidasan et al. 2021; H. Zhang et al. 2018). The goal of OE is to expose the model to OOD data during training, in order to improve its ability to recognize and classify OOD data at test time. This has been shown to improve OOD detection performance across a wide range of benchmarks.

However, in certain applications such as medical imaging, it can be difficult to obtain representative datasets for OE. As a result, there has been a recent focus on OE-free methods, such as Generalized-ODIN (Hsu et al. 2020) which uses ID data samples to tune the required hyper-parameters for detection instead of traditional ODIN and Virtual Outlier Synthesis (Du et al. 2022) which exposes the model to synthetically generated latent space outliers.

2.2.5 Evaluation Metrics

There are several evaluation metrics commonly used to measure the performance of out-of-distribution (OOD) detection methods. These include,

- *Area under the receiver operating characteristic curve (AUROC)*:This metric measures the trade-off between true positive rate (TPR) and false positive rate (FPR) at different classification thresholds. A model with a high AUROC score has a low FPR and a high TPR, indicating that it is able to accurately identify both in-distribution (ID) and OOD data.
- *Area under the precision-recall curve (AUPRC)*: This metric measures the trade-off between precision (the fraction of correct predictions among all predictions) and recall (the fraction of correct predictions among all positive examples) at

different classification thresholds. A model with a high AUPRC score has a high precision and a high recall, indicating that it is able to accurately identify both ID and OOD data.

- *False positive rate at 95% true positive rate (FPR95)* This metric measures the FPR when the TPR is at 95%. It can be used to compare the OOD detection performance of different models at a fixed TPR.

It is important to choose the appropriate evaluation metric depending on the specific goals and requirements of the OOD detection task.

2.3 Uncertainty Quantification in Deep Neural Networks

In statistics, uncertainty quantification (UQ) is a powerful tool that studies and identifies the impact of different error sources on model prediction. Sources of error in a model training pipeline include data sampling, model selection, the inherent randomness as well as their complex interactions. By accounting for the sources of potential errors, UQ allows the rigorous characterization of deep models by allowing model predictions to be represented as a distribution in lieu of conventional point estimates. UQ aids in measuring how well a model reflects the physical reality as well as whether the model adheres to the human understanding of an underlying process supporting actionability.

Broadly, a rigorous characterization of ML and DNN tools will help in:

- Building reliable models that are consistent between predictions and our understanding of the process
- Avoid model overconfidence when dealing with inputs from different test distributions

- Identify improperly sampled data regimes
- Design human-in-the-loop systems

Consequently, there is a strong need to estimate prediction and representation uncertainties in deep models to understand model predictions. These uncertainty estimates enable the safe and reliable adoption of ML models and find practical use in applications such as anomaly detection (Anirudh 2021), adversarial defence and robustness under distribution shifts (Lakshminarayanan, Pritzel, and Blundell 2017) etc. The error sources in the training pipeline can be broadly grouped into two classes of uncertainties namely (i)*aleatoric* and (ii) *epistemic* uncertainties.

2.3.0.1 Aleatoric Uncertainty

Aleatoric uncertainty also referred to as statistical uncertainty is due to the stochastic properties of the data or the data collection mechanism. This uncertainty cannot cannot be further reduced even under the limit of infinite data. Derived from the Greek word *alea* referring to dice games, aleatoric uncertainty is used to represent phenomena such as thermal noise in a physical process, human errors etc. In the context of ML, strategies such as data augmentation/providing different views of the data while training can be used to minimize this uncertainty.

2.3.0.2 Epistemic Uncertainty

Epistemic uncertainty arises due to the lack of knowledge of the underlying process - how to effectively sample data from different data regimes, model selection etc. In terms of ML, this corresponds to a scenario where our model parameters and hence

predictions are poorly estimated. Such uncertainties can be completely eliminated under the availability of infinite data and appropriate knowledge about the underlying model.

2.3.1 Existing Approaches for Estimating Uncertainties in Deep Neural Networks

Bayesian methods (Neal 2012) which aim to estimate the predictive posterior distributions are among the commonly adopted uncertainty estimation strategies. While such approaches can best approximate the posterior, they are computational bottlenecks when scaled to larger datasets. Variational methods such as Monte Carlo Dropout (Gal and Ghahramani 2016) is a scalable alternative to Bayesian methods in that it approximates the posterior distribution on the weights via dropout to estimate uncertainties. However, it is well known that dropout tends to over estimate uncertainties due to sampling the parameter space in regions that are potentially far away from the local minima. Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) showed a simple way to obtain accurate uncertainties in a task agnostic and model agnostic fashion. While extremely accurate and currently one of the best uncertainty estimation techniques, the authors of (Ovadia et al. 2019) showed that it requires training several models with different random initializations which can become a computational bottleneck when training deep networks. As a result, there have been several methods that seek to obtain accurate uncertainties from just a single model, typically a DNN. Deterministic Uncertainty Quantification (Van Amersfoort et al. 2020) uses a kernel distance to a set of class-specific centroids defined in the feature of a deep network as the measure for uncertainty. Other techniques include Direct Epistemic Uncertainty Prediction (Jain et al. 2021) which trains an explicit

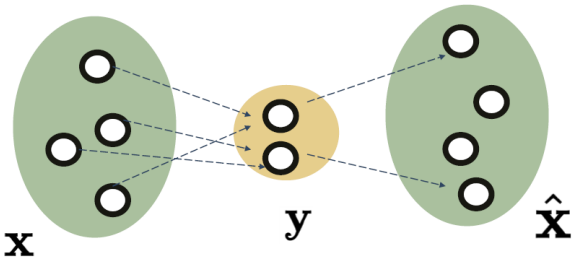


Figure 6. Forward and Inverse Mappings Between Inputs and Outputs. For Solving Inverse Problems, We Are Interested in Estimating the Parameters That Best Explain a given Observation. Since Such Mappings Are Seldom Bijective, the Reliable Estimation of the Inputs Can Become Significantly Challenging.

epistemic uncertainty estimator for a pre-trained model, which can also be used efficiently in a task-agnostic manner.

2.4 Inverse Problems

In a wide-range of applications in science and engineering, one often faces the need to learn complex mappings between independent parameters and dependent/measured quantities, i.e. the *forward* and *inverse* mappings. As illustrated in Fig.6, the forward process can be defined as $\mathcal{F} : \mathbf{x} \mapsto y$ where $\mathbf{x}$ represents a set input parameters. $\mathcal{F}$ is the forward process which transforms $\mathbf{x}$ to an observation y. The forward function $\mathcal{F}$ may or may not be known *a*priori and valid assumptions are made to represent the same. For instance, $\mathcal{F}$ can be a simple, analytical function such a known noise process or it can assume arbitrarily complex forms such as a differentiable neural network or non-analytical simulators. Mathematically, an inverse problem is formulated as,

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}}\ \mathcal{L}\Big(\mathbf{y}, \mathcal{F}(\mathbf{x})\Big) + \lambda\mathcal{R}(\mathbf{x}), \tag{2.6}$$

where $\mathcal{L}(.)$is a suitable modality dependent objective function that penalizes the optimization when the prediction does not match the given observations. Here, $\mathcal{R}(.)$

is a regularizer or prior adopted to control the overall inversion. When the mappings between the input and output are bijective, Equation 2.6 is well-defined and tractable. However, in real world scenarios, the mappings are seldom bijective and ill-posed. Hence the reliable estimation of the input parameters and hence the exploration of the parameter space can become significantly challenging. The inverse optimization can converge to trivial or sub-optimal solutions that satisfy the objective function $\mathcal{L}$. The regularizer plays a key role in controlling the inversion by placing plausible assumptions on the nature of the estimated solutions. In other words, it reduces the search space for this ill-posed optimization allowing optimal convergence. However, the choice of the prior plays a crucial role in determining the overall quality of the estimates. This is because, the true statistical properties and structure of the solutions can be different from the properties imposed by the prior.

2.4.1 Priors in Inverse Problems

Priors which place meaningful assumptions on the structure on the inferred solutions to regularize the inversion can be broadly classified into three categories namely (i) Classical Priors, (ii) Generative Priors and (iii) Model based priors.

2.4.1.1 Classical Priors

These priors are based on imposing suitable mathematical constraints on the structural properties of the inferred solutions. A broad group of classical priors include the class of matrix factorization methods conventionally used for source separation. For example in ICA (Hyvärinen 1999), we enforce the assumptions of

non-Gaussianity and statistical independence on the estimated inputs. On the other hand, PCA (Bishop and Nasrabadi 2006) enforces statistical independence by linear projection onto mutually orthogonal subspaces. KernelPCA (Mika et al. 1999) induces the same prior in a reproducing kernel Hilbert space. Another popular approach is Non-negative matrix factorization (NMF), which places a non-negativity prior on the estimated basis matrices (Févotte, Vincent, and Ozerov 2018). A sparsity prior (ℓ_1) (Virtanen 2003) placed either in the observed domain or in the expansion via an appropriate basis set or a dictionary are also powerful priors. Classical priors adopted for ill-posed image restoration tasks include total variation and the ℓ_2 norm. Total variation (Mahendran and Vedaldi 2015) encourages images to contain piece-wise constant patches while the ℓ_2 *norm* regularizes the range and energy of the image to remain within a given interval.

2.4.1.2 Generative Priors

A more recent class of inverse problems (Shah and Hegde 2018; Anirudh et al. 2020) have relied on priors defined via generative models, e.g. Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). GANs learn parameterized non-linear distributions $p(X; \mathbf{z})$ from a sufficient amount of unlabeled data X (Donahue, McAuley, and Puckette 2019; Radford, Metz, and Chintala 2015), where $\mathbf{z}$ denotes the latent variables of the model. In addition to readily sampling from trained GAN models, they also serve as an effective prior for X. In its most general form, when one attempts to solve the inversion of recovering the original data $\mathbf{x}$ from its corrupted version $\tilde{\mathbf{x}}$, one can maximize the posterior distribution $p(X = \mathbf{x}|\tilde{\mathbf{x}}; \mathbf{z})$ by searching in the latent space of a pre-trained GAN. Since this posterior distribution cannot be expressed

analytically, in practice, iterative approaches such as *Projected Gradient Descent* (PGD) (Anirudh et al. 2020) are used in practice.

2.4.1.3 Model Based Priors / Structural Priors

Recent advances in deep neural network design have shown that the structure of certain carefully chosen networks have the innate capability to effectively regularize or behave as a prior to solve ill-posed inverse problems. These networks essentially capture the underlying statistics of data, independent of the task-specific training. These *structural priors* have produced state-of-the-art performance in inverse imaging problems (Ulyanov, Vedaldi, and Lempitsky 2018) as well as for audio source separation (Tian, Xu, and Li 2019). Popular examples for model prior architectures include an untrained, randomly initialized U-Net (Ronneberger, Fischer, and Brox 2015) and Implicit Neural Representations (Sitzmann et al. 2020; Tancik et al. 2020) with Fourier feature based positional encoding and sinusoidal activations. The main reason behind the regularization capability of such networks is that these networks offer the innate capacity to reject noise and accept natural image statistics. In comparison with the other class of priors, model based priors re-parameterize the estimated image in terms of the learnable weights of the prior architecture. This allows the optimization to search over a high dimensional space to infer meaningful solutions. Model based priors can be directly trained during inference time with a single sample that needs to be restored.

2.5 Summary

In this chapter, we provided an overview of the foundational concepts related to the problems addressed in this dissertation. We started by discussing the importance of accurately characterizing confidence in deep learning models, including model calibration and the advantages and disadvantages of different metrics for representing confidence. We also introduced the problem of out-of-distribution detection, describing the current strategies and scoring functions used, as well as the various types of out-of-distribution data that a detector needs to handle. We then highlighted the importance of uncertainty quantification for accurately characterizing confidence and making informed decisions, and described the different sources of uncertainty and the current methods for estimating them. Finally, we covered the topic of inverse problems and the existing approaches for addressing them.

Chapter 3

DESIGNING DIRECT ERROR PREDICTORS FOR CHARACTERIZING MODEL CONFIDENCE

In this chapter, we consider the problem of effectively characterizing confidence of deep deterministic models. In general, confidences of DNNs have been found to be poorly calibrated (Guo et al. 2017) not meeting the expectation of the end-user or evidence provided by the training data. We identify gaps in the existing measures adopted for estimating confidence and find that such metrics are dependent solely on the training dynamics without considering the underlying data and model stochasticities. We develop a novel strategy to train a surrogate to estimate loss of the underlying model which is a reliable metric for characterizing confidence. Interestingly, we find that such a surrogate offers useful properties that can be leveraged to perform model introspection via feature importances and counterfactual explanations. We observe that such surrogates are robust to mild distribution shifts to which even the underlying model generalizes to.

3.1 Problem Setup

Deep learning methods are routinely used in state-of-the-art AI models for a wide variety of applications (Devlin et al. 2019; Karras, Laine, and Aila 2019) handling a diverse set of data modalities. In particular, over the last few years, we have witnessed advances to the use of AI even in prescribing potential actions that need to be taken even in critical applications such as healthcare (Hosny et al. 2018; Young et al. 2020).

A key step towards promoting the adoption of these tools in practice is to ensure that the models behave predictably, and do not provide unintended generalization to regimes where the training data provide no meaningful evidence. To this end, we focus on characterizing the behavior of DNNs both in terms of generalization as well as reliably detecting distribution shifts (Vivek Narayanaswamy et al. 2021; Cao, Huang, et al. 2020).

It is important to note that a model that produces high accuracy on the original data distribution is not always guaranteed to be effective at detecting distribution shifts. On the other hand, models that are overly sensitive to even mild distribution shifts can provide inferior generalization performance. This clearly emphasizes the need for learning strategies that can effectively regulate the model predictions, such that the model is well-calibrated to detect such shifts (and can defer from making predictions), while not trading off performance.

Popular approaches adopted for characterizing model predictions include off-the-shelf metrics such as the explicitly defined maximum softmax probability (Guo et al. 2017). However, such scores extensively rely on the training dynamics, does not take into account the inherent data and model stochasticities and hence are found to be poorly calibrated in practice. It has been well demonstrated that DNNs produce confident predictions even for data not evidenced by the training distribution. Other measures include (i) predictive entropy (Lakshminarayanan, Pritzel, and Blundell 2017) which characterizes the randomness in predictions as a proxy for model confidence and, (ii) energy (Liu et al. 2020) that maps the input probability density to an unconstrained scalar. While these measures are significantly superior to the softmax interpretation of confidences, such metrics are not calibrated well enough under distribution shifts.

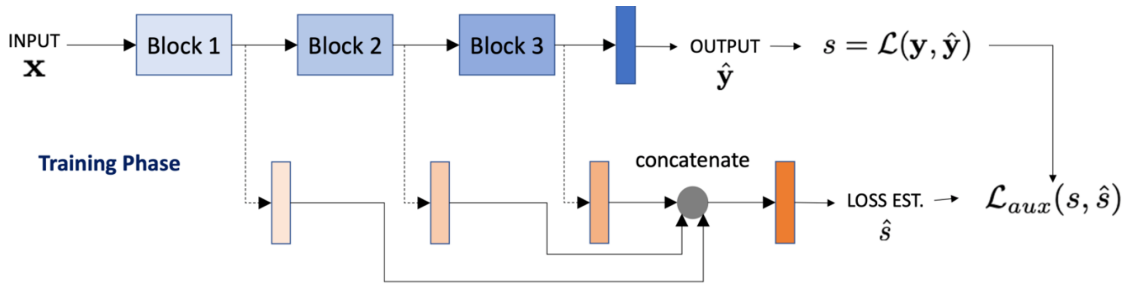


Figure 7. An Overview of Our Approach. We Propose to Jointly Train a Direct Error Predictor (DEP) Alongside the Classifier in Order to Obtain the Prediction Uncertainties. This Joint Training Process in Turn Regularizes the Classifier Model Training and Helps Produce Well-calibrated Predictions.

Post-hoc calibration strategies are popular approach for ensuring that the distribution of model prediction probabilities matches the true distribution from the observed data. For example, Platt scaling (Guo et al. 2017) and isotonic regression (Kuleshov, Fenner, and Ermon 2018) based on a validation dataset. Approaches that temper model confidences based on explicit uncertainty estimators have also been proposed (Seo, Seo, and Han 2019; Thiagarajan, Venkatesh, Sattigeri, et al. 2020). While the effectiveness of the former approach relies heavily on the choice of the validation dataset, epistemic uncertainty estimators used in the latter class of approaches are also found to be poorly calibrated (Kuleshov, Fenner, and Ermon 2018). In general, epistemic uncertainty refers to the lack of knowledge and could be eliminated with sufficient data – such an optimal learner is referred to as the Bayes optimal predictor (Bishop and Nasrabadi 2006; Jain et al. 2021). The irreducible error associated with the Bayes optimal predictor is the aleatoric uncertainty.

In this work, we design a *direct error predictor* (DEP) as a surrogate to predict the generalization error which is a reliable metric for model failure. Recent studies on active learning and explainable AI (Yoo and Kweon 2019; JJ. Thiagarajan et al. 2021) have found that such an error predictor can capture the inherent model

uncertainties. While a DEP can be learned for any pre-trained predictor (Jain et al. 2021), we make an interesting finding that jointly training the DEP alongside a predictor produces calibrated uncertainties. Our contributions can be summarized as follows: (i) We perform confidence characterization based on a direct error predictor which is trained alongside the underlying model using novel training objectives namely contrastive loss and dropout calibration loss; (ii) Using a challenging skin lesion detection dataset (Codella et al. 2018), we demonstrate improved model generalization to in-distribution data using standard classification metrics; (iii) Using a state-of-the-art OOD detector (Liang, Li, and Srikant 2018), we show that our predictor regularized by DEP achieves significant improvements in detecting out of distribution samples even on new classes not seen during training; (iv) We make a crucial observation that such DEPs can effectively be utilized for introspecting models via feature importances and counterfactual explanations. We elaborate upon these applications in separate chapters where we provide extensive empirical studies and comparisons with state-of-the art baselines for the same.

3.2 Predictive Model Design with Direct Error Prediction

We consider the setup where we build a predictive model $\mathcal{F}(\Theta)$ which takes as input a sample $\mathbf{x} \in \mathbb{R}^d$ with d features and produces the output $\hat{\mathbf{y}} \in \mathbb{R}^k$ of dimensionality k. Given a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, we optimize for the parameters Θ using the loss function $\mathcal{L} : \mathbf{y} \times \hat{\mathbf{y}} \rightarrow s$, where $s \in \mathbb{R}$. In other words, $\mathcal{L}$ measures the discrepancy between the true and predicted outputs using a pre-specified error metric. Examples include categorical cross-entropy for classification or mean-squared error for regression.

Our approach requires the training of an auxiliary network $\mathcal{G}(\Phi; \Theta)$ that takes the

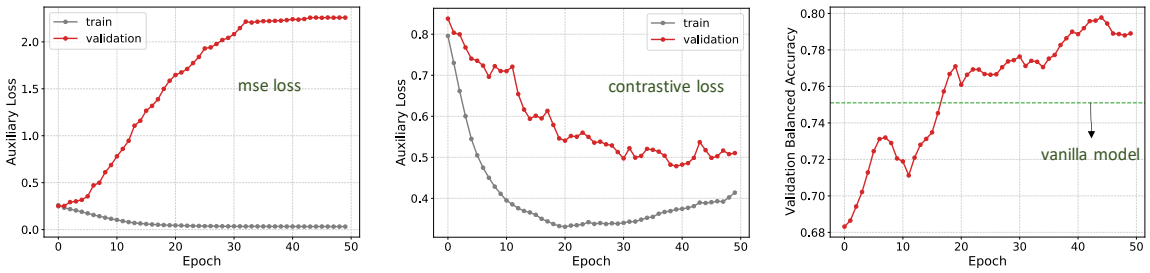


Figure 8. Learning Behavior of the Joint Training Process on the ISIC 2019 Skin Lesion Detection Dataset. Using a Contrastive Objective for the Auxiliary Loss Makes the DEP Generalize to the Validation Data (Middle) and Regularizes the Predictor to Improve Its Accuracy (Right). In Contrast, Implementing the Auxiliary Loss as the Standard MSE Objective Provides Very Poor Generalization to Unseen Data (Left).

same input $\mathbf{x}$ and produces the output $\hat{s} \approx \mathcal{L}(\mathbf{y}, \mathcal{F}(\mathbf{x}))$. The objective of this network is to directly estimate the fidelity for the prediction that $\mathcal{F}$ makes for $\mathbf{x}$. Note that, the loss estimates implicitly provide information about the inherent uncertainties; for example, in (Ash et al. 2020), the gradients of loss estimates have been used to capture the model uncertainties. We define the auxiliary objective $\mathcal{L}_{aux} : s \times \hat{s} \rightarrow \mathbb{R}$, in order to train the parameters Φ of model $\mathcal{G}$. As showed in Figure 7(top), the loss estimator $\mathcal{G}$ uses the latent representations from different stages of $\mathcal{F}$ (e.g., every layer in the case of an FCN or every convolutional block in a CNN) to estimate $\hat{s}$. We use a linear layer along with non-linear activation (ReLU in our experiments) to transform each of the latent representations from $\mathcal{F}$ and they are finally concatenated to predict the loss. During training, the gradients from both the losses are used to update the parameters Θ of model $\mathcal{F}$.

3.2.1 Learning Objectives

Since we require the direct error predictor to be consistent with the underlying predictor, the choice of the loss function $\mathcal{L}_{aux}$ to train the DEP is crucial. An important property that is expected from DEPs is that it preserves the ordering of samples (based on their losses), even if the original scale is discarded. For example, one can use the mean squared error (MSE) objective. However, as showed in Figure 8 (left), DEPs trained with the MSE objective do not generalize to new data, since the DEP converges to produce an average loss value for all data. Hence, we explore the following two loss functions to train DEPs which aim to preserve the ordering of samples based on their corresponding losses from $\mathcal{F}$.

3.2.1.1 Contrastive Training

This is a widely adopted strategy when relative ordering of samples needs to be preserved. Given the loss values $\{s_i, s_j\}$ for a pair of samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ in a mini-batch, we adopt an objective similar to (Yoo and Kweon 2019), which ensures that the sign of the difference $(s_i - s_j)$ is preserved in the corresponding loss estimates $(\hat{s}_i - \hat{s}_j)$. Formally, we use the following contrastive loss:

$$\mathcal{L}_{aux} = \sum_{(i,j)} \max\Big(0, -\mathbb{I}(s_i, s_j).(\hat{s}_i - \hat{s}_j) + \gamma\Big), \tag{3.1}$$

$$\text{where } \mathbb{I}(s_i, s_j) = \begin{cases} 1 & \text{if } s_i > s_j, \\ -1 & \text{otherwise.} \end{cases}$$

Note, when the sign of $s_i - s_j$ is positive, we assign a non-zero penalty if the estimates $\hat{s}_j > \hat{s}_i$, i.e., there is a disagreement in the ranking of samples. Here, γ is an optional

margin hyper-parameter. As showed in Figure 8 (right), DEPs trained with the contrastive objective generalizes well to the validation data.

3.2.1.2 Dropout Calibration

In this formulation, we utilize prediction intervals from the model $\mathcal{F}$ and adjust the loss estimates from $\mathcal{G}$ using an interval calibration objective. The notion of interval calibration comes from the uncertainty quantification literature and is used to evaluate uncertainty estimates in continuous-valued regression problems (Thiagarajan, Venkatesh, Sattigeri, et al. 2020). In particular, we consider the epistemic uncertainties estimated using Monte Carlo dropout (Gal and Ghahramani 2016) to define the prediction interval $[\mu_{s_i} - \sigma_{s_i}, \mu_{s_i} + \sigma_{s_i}]$ for a sample $\mathbf{x}_i$. More specifically, we perform T independent forward passes with $\mathcal{F}$ to compute the mean μ_{s_i} and standard deviation σ_{s_i}. For the loss estimator $\mathcal{G}$, we use the latent representations averaged across T passes (for every block in Figure 7(top)) to obtain the estimate $\hat{s}_i$. Finally, we use a hinge loss objective to calibrate the estimates:

$$\mathcal{L}_{aux} = \sum_i \max\left(0, \hat{s}_i - (\mu_{s_i} + \sigma_{s_i}) + \xi\right) \tag{3.2}$$

$$+ \max\left(0, (\mu_{s_i} - \sigma_{s_i}) - \hat{s}_i + \xi\right) \tag{3.3}$$

Here, ξ is the optional margin parameter and the objective encourages the estimates $\hat{s}_i$ to lie in the prediction interval for s from the model $\mathcal{F}$.

The overall objective for the joint optimization of the predictor and DEP is given by

$$\mathcal{L}_{total} = \mathcal{L} + \lambda \mathcal{L}_{aux}. \tag{3.4}$$

Interestingly, as showed in Figure 8, using the contrastive objective produces general-

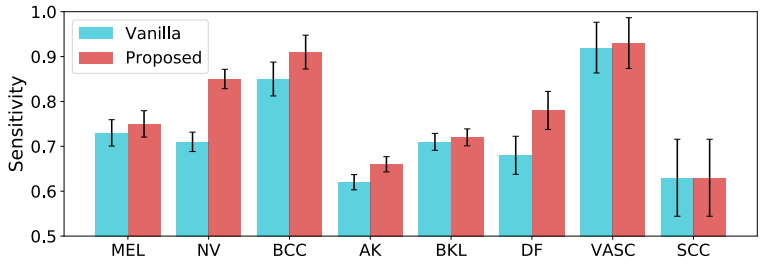
izable DEPs and also effectively regularizes the predictor model training (indicated by significant improvement in the validation accuracy over an unregularized model).

3.3 Impact of Jointly Training Classifiers with Direct Error Predictors on Generalization

In this experiment, we build predictor models with and without an additional DEP module (Fig. 7), and present comparisons on the ISIC 2019 (Codella et al. 2019) dataset.

3.3.1 Experiment Setup

We learned the predictor models using the ISIC 2019 dataset (Task1) – we used $90-10$ stratified random splits for train/validation and report performance from 3 independent trials. We adopted standard data pre-processing steps used for this data (Gessert et al. 2020) – all images were center cropped to 0.85 times its original width, adjusted for color constancy and resized to $224 \times 224 \times 3$. Further, we performed data augmentation (horizontal/vertical flips, jitter, rotation and translation). For all experiments, we adopted a VGG-16 network (pre-trained on imagenet) and fine-tuned it for 50 epochs. The DEP utilized the activations of the intermediate convolutional layers 4, 7, 10 and 13 from the predictor followed by affine transformations using linear layers of 128 units each. These individual representations are then concatenated and transformed using a final linear layer to obtain the loss estimate for that sample. For model training, we used the following hyper-parameters: learning rate of $1e-4$ reduced by a factor of 0.5 every 10 epochs, Adam optimizer with 0.9 momentum and



Balanced Accuracy (%)

Vanilla	Proposed
73.1 +/- 0.6	**77.9 +/- 1.5**

Figure 9. Performance Comparison of the Vanilla and Our Proposed Models Using Sensitivity and Balanced Accuracy Metrics (Results Averaged Across 3 Independent Trials) on the ISIC 2019 Lesion Dataset.

$5e-4$ weight decay, dropout of 0.4, and batch size of 64. In addition, to handle the class imbalances and perform well on undersampled classes (e.g. AK, DF, BCC), we used weighted random sampling and a weighted cross-entropy loss. Following common practice, we compared the models using the class-specific sensitivity and balanced accuracy scores.

3.3.2 Results

As showed in Figure 9, our approach convincingly outperforms the vanilla model in classifying in-distribution data. Jointly training the DEP alongside the predictor (i) allows training with higher learning rates (faster convergence), (ii) achieves higher performance on minority classes (BCC, AK, DF), while also improving performance for densely sampled classes (NV, MEL), and (iii) produces $\sim 5\%$ improvement over the vanilla model in terms of balanced accuracy score (77.9% as opposed to 73.1% of the vanilla model).

3.4 Impact of Jointly Training Classifiers with Direct Error Predictors on Out-of-Distribution Detection

In this section, we setup an experiment to study and evaluate the effectiveness of the DEP-regularized predictor in detecting out-of-distribution data. We begin by describing the training setup, the OOD detector implemented following (Liang, Li, and Srikant 2018), and the metrics utilized to for evaluation.

3.4.1 Experiment Setup

In this study, we utilized 5 out of the 8 lesion image categories from the ISIC 2019 training dataset, namely MEL, NV, BCC, AK and DF, to define the in-distribution data. This selection amounts to a total of $21,826$ samples. We performed a $90-10$ stratified random split of images from these 5 classes to train and validate the models. We used the same architectures as the previous experiment to implement the predictor and DEP. Note, we used the ADAM optimizer with a learning rate of $1e-4$ and trained the models for 50 epochs. For comparison, we also trained a vanilla model under the same experiment settings.

3.4.2 ODIN Detector

In order to evaluate the model reliability in rejecting samples (defer to make predictions) when presented images from distinctly different distributions, we adopt ODIN (Out of DIstribution detector in Neural networks) (Liang, Li, and Srikant 2018), a popular approach for detecting OOD data based upon the calibrated confidence

(a) B-Box (b) B-Box-70 (c) Imagenet (d) NCT (e) Clin Skin (f) Derm Skin

Figure 10. Examples from the OOD Datasets Used for Our Experiments.

(softmax) scores. In particular, ODIN employs temperature scaling to the softmax probabilities and applies controlled input image perturbations to enlarge the confidence score gap between the in-distribution and out-of-distribution data. For an input sample $\mathbf{x}$, ODIN first computes the confidence score from temperature (T) scaled softmax probabilities:

$$S(\mathbf{x}, T) = \max_{c} \frac{\exp(\mathcal{F}_c(\mathbf{x})/T)}{\sum_{k=1}^{K} \exp(\mathcal{F}_k(\mathbf{x})/T)} \tag{3.5}$$

where $\mathcal{F}_c(\mathbf{x})$ is the predictive model output for the c^{th} class in a K-way classification problem. In addition to temperature scaling, ODIN also systematically perturbs the input by a factor of η in the direction of the gradient of the loss w.r.t $\mathbf{x}$, in order to improve the softmax probabilities. Formally,

$$\hat{\mathbf{x}} = \mathbf{x} - \eta \ \text{sign}(-\nabla_{\mathbf{x}} \log(S(\mathbf{x}, T)) \tag{3.6}$$

Finally, the confidence score from the perturbed image is used to determine if x is an inlier sample – based on a user-specified threshold γ.

3.4.3 OOD Datasets

In order to demonstrate the behavior of the predictor under a wide variety of unknown data regimes, we consider the following benchmark datasets, following (Pacheco

Table 1. Evaluating the OOD Detection Performance of the ODIN Detector Using Both the Vanilla and Proposed Methods. The In-distribution Data Includes Validation Set Samples Belonging to the 5 Classes. The Hyper-parameters for ODIN are Fine Tuned on the NCT Dataset and Evaluated on the Remaining Datasets.

OOD Dataset	**Metrics in %**				
	FPR@TPR95	**DTERR**	**AUROC**	**AUPR-In**	**AUPR-Out**
	Vanilla/Proposed				
NCT	51.41/**17.93**	21.44/**10.19**	87.27/**96.43**	91.41/**97.73**	81.71/**94.43**
B-Box	36.43/**5.16**	17.84/**4.91**	90.74/**98.74**	90.91/**98.71**	91.15/**98.71**
B-Box-70	**0.0**/**0.0**	**0.1**/**0.1**	**100.0**/**100.0**	**99.97**/**99.97**	**99.98**/**99.98**
ImageNet	71.05/**47.35**	28.7/**20.92**	79.03/**87.57**	81.68/**87.74**	76.38/**86.84**
Clin-Skin	65.56/**35.13**	25.0/**14.11**	82.58/**92.15**	93.37/**96.71**	61.07/**79.09**
Derm-Skin	56.04/**21.6**	21.6/**10.38**	87.14/**95.69**	91.22/**97.2**	82.59/**92.14**

et al. 2020). In particular, we consider (i) *B-Box*: 2025 skin lesion images corrupted by a black bounding box on the lesion region; (ii) *B-Box-70*: 2454 skin lesion images with black bounding boxes that mask $\sim 70\%$ of the lesion region; (iii) *ImageNet* (Deng et al. 2009): - 3000 images randomly chosen from the ImageNet database; (iv) *NCT*: 1350 histopathology images of human colorectal cancer acquired from (Kather et al. 2019) (v) *Clin-Skin*: 723 clinical images of healthy skin; (vi) *Derm-Skin*: 1565 dermoscopy images of skin obtained by randomly cropping patches in the ISIC 2019 dataset. Figure 10 shows sample images from each of these datasets.

3.4.4 Evaluation Metrics

We adopt these widely adopted metrics to quantify the OOD detection performance: (i) **False Positive Rate** @95**% True Positive Rate**(FPR@TPR95): Probability that an OOD sample is misclassified as an in-distribution sample when the TPR is as high as 95 %; (ii) **Detection Error**(DTERR): Minimum probability of mis-detecting

an inlier sample as an OOD sample over all possible thresholds; (iii) **AUROC**: Area Under the Receiver Operator Characteristic curve (TPR vs FPR) is a threshold independent metric which reflects the probability that an in-distribution image is assigned a higher confidence over the OOD sample; (iv) **AUPR-In** and **AUPR-Out**: Area under the Precision-Recall curve where the in distribution and the OOD samples are considered as positives respectively.

3.4.5 OOD Detection Performance

In Table 1, we report the OOD detection performance of our approach against the baseline vanilla predictor over a variety of datasets characterized by apparent distribution changes. The hyper-parameters T and η were fine tuned for both the models using the NCT dataset and evaluated on the remaining datasets. It must be noted that while using the ODIN detector, we utilized the validation split of the 5 class ISIC 2019 dataset as the inlier distribution and the other listed datasets as OOD to compute the metrics. It can be observed from Table 1 that our model, which is effectively regularized by the loss estimator, significantly outperforms the baseline, thus implying improved model reliability under data regime changes. While it is more challenging to detect OOD samples which exhibit semantic similarities with the in-distribution data (Cao, Huang, et al. 2020), e.g., *Clin-Skin* and *Derm-Skin*, we find that by coupling the uncertainty estimator during training, we can still effectively detect these shifts with improved specificity.

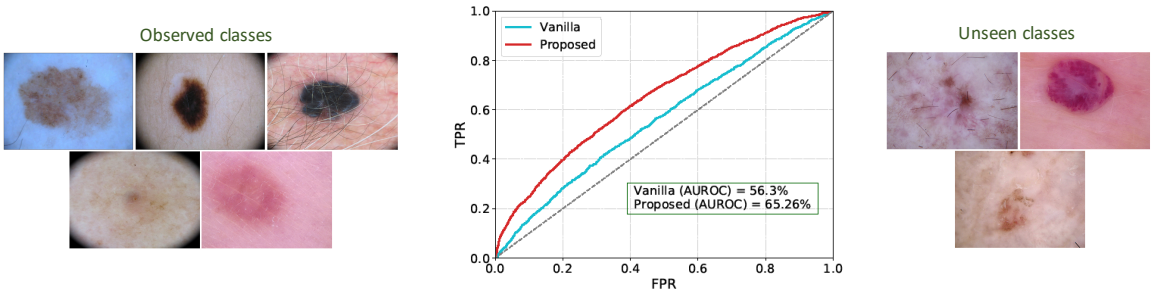


Figure 11. Performance of Our Approach in Detecting New Classes Unseen During Training.

3.4.6 Detecting New Classes

In this experiment, we evaluate the ability of our approach in detecting classes of data that are unseen during model training. Similar to the previous experiment, we utilized both the vanilla and the proposed models trained on the 5 class ISIC 2019 dataset. Subsequently, we introduced images from the 3 remaining classes (BKL, VASC and SCC) as novel data at test time. We expect a reliable model to effectively detect these samples as OOD, thus enabling us defer from making an incorrect diagnosis. We find that the task of detecting new classes is significantly more challenging due to the less apparent semantic discrepancies between the observed and the unseen classes. From the results in Figure 11, we notice that our approach still outperforms standard deep models (in terms of the AUROC metric – Vanilla: 56.3%, Our Approach : 65.26%) in detecting samples from unseen classes. This further emphasizes the value of DEPs in controlling unintended generalization of predictive models.

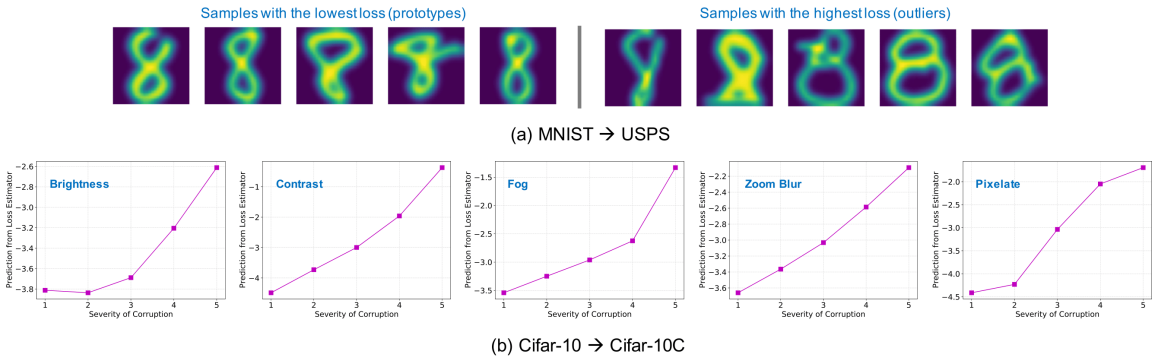


Figure 12. Effectiveness of DEP $\mathcal{G}$ in Detecting Distribution Shifts, Even Though the Shifts Are Not Known During Training. In the MNIST-USPS Case, it Attributes Non-typical Writing Styles from the USPS Dataset, That Are Not Found in MNIST, with High Loss Values. Similarly, in the Case of CIFAR-10C, the Loss Estimates from $\mathcal{G}$, Averaged Across 500 Test Samples, Monotonically Grows as the Severity of the Corruption Increases.

3.5 Properties of DEP

3.5.1 Detecting Distribution Shifts and Identifying Prototypes

At the core of our approach is the pre-trained DEP, which serves as a reliable failure indicator. Consequently, the robustness of DEP directly relies on how well it can generalize under distribution shifts. We investigate the empirical behavior of DEP using (i) MNIST-USPS and (ii) CIFAR-10 to CIFAR10-C benchmarks. In both cases, we train the predictor and DEP using the original data (MNIST, CIFAR-10) and evaluate on the shifted data. In Figure 12(a), we show USPS images from class 8 with the lowest (in-distribution) and highest (out-distribution) loss estimates. While the former resemble the prototypical examples, the latter contains uncommon writing styles not found in the MNIST dataset. In case of CIFAR-10-C, we show the loss estimates for 5 different natural image corruptions (averaged across 500 examples).

We observe a monotonic increase in the average loss estimates as the severity of the corruptions grow, thus demonstrating the ability of DEP to detect distribution shifts.

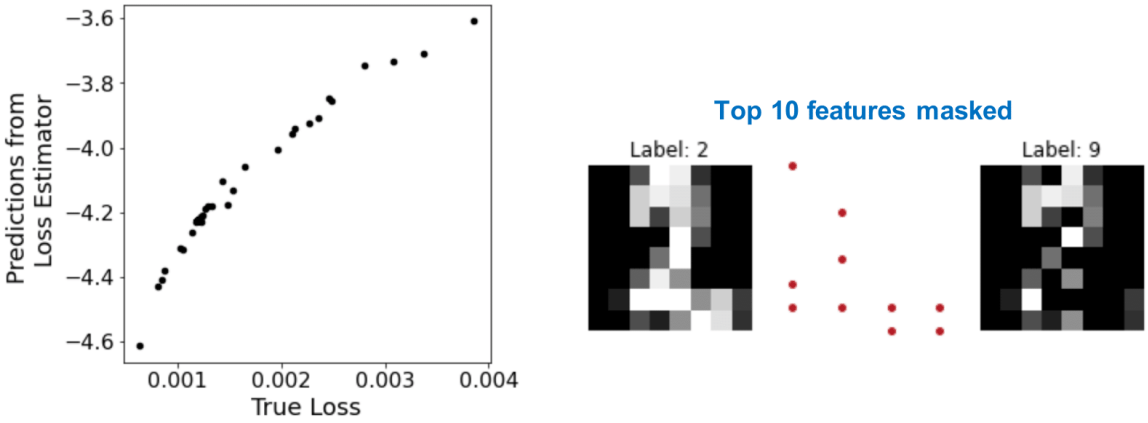


Figure 13. Demonstrating the Pixel-sensitivity Property of DEP on the UCI Handwritten Digits Dataset. Here, We Show an Example Where DEP Was Trained Using the Contrastive Loss. For This Test Sample, the Ranking Obtained Using the Estimated Loss Agrees with That from the True Loss (Known Ground Truth). When We Mask the Top 10 Features from DEP and as Expected, There is a Change in the Model Prediction.

3.5.2 Sensitivity to Pixel/Feature Level Manipulations

For this study, we consider the UCI handwritten digits dataset (Dua and Graff 2017) comprised of 8×8 grayscale images. In Figure 13, we show predictions from our DEP (contrastive training) for a test image, when each of the 64 pixels were masked (replaced with zero). We find that, though the scale of the loss function is discarded, the ordering of the features is well preserved. We also illustrate the explanation obtained by masking the top 10 features identified. The observed changes in the prediction (from class 2 to class 9) is intuitive and demonstrates that DEP is sensitive to pixel-level variations which can be adopted for supporting model introspection.

3.6 Utility of Direct Error Predictors for Model Introspection

3.6.1 Feature Importance Explanations

With increasing reliance on the outcomes of black-box models in critical applications, explainability tools that do not require access to the model internals are often used to enable humans understand and trust these models. In particular, the class of methods that can reveal the influence of input features on the predicted outputs are useful. We find that DEP which is consistent with the underlying model can be effectively repurposed to estimate the importance of every feature in an input towards the model prediction. The DEP can directly generate post-hoc explanations by measuring the influence of input features (masking) on the model output using a Granger causal objective defined on the loss estimates. We find that training such DEPs jointly with the classifier can accurately estimate the feature importance scores even under complex distribution shifts, without any additional re-training unlike existing methods (Lakkaraju, Arsov, and Bastani 2020). The properties of the DEP discussed above enable their usage in model introspection. Chapter 4 provides an extensive analysis on the usage of DEPs for such explanations along with comparisons over multiple datasets, architectures and baseline methods.

3.6.2 Counterfactual Explanations

Counterfactual (CF) explanation techniques are those that synthesize small, interpretable changes to a given image while producing desired changes in the model prediction. CF explanations provide more flexibility and interpretability over con-

ventional techniques, such as feature importance estimation (Selvaraju et al. 2017; Lakkaraju, Arsov, and Bastani 2020; Shrikumar, Greenside, and Kundaje 2017; Ribeiro, Singh, and Guestrin 2016; Ribeiro, Singh, and Guestrin 2018) and are becoming exceedingly popular for introspecting black-box models. In this work, we focus on the case where we have access only to the trained deep classifier and not the actual training data. The key requirements for CF explanations to be useful are that they are required to contain discernible changes (for easy interpretability) while being realistic (consistency to the data manifold). Synthesizing CFs that satisfy the requirements is usually formulated as an inverse problem regularized by suitable priors. Existing strategies to perform such an inversion without any access to training data utilize weak priors such as ℓ_2 or total variation lead to poor quality explanations. Although, the quality of CFs can be improved using strong image priors (Ulyanov, Vedaldi, and Lempitsky 2018), such priors can easily overfit when the data to be explained in out-of-distribution. In such cases, we identify that the pre-trained DEPs can be operated as a self-calibrated hypothesis test consistent with the underlying model to provide meaningful gradients to the image generation process to synthesize inlier manages. We find that DEP effectively re-projects any OOD data onto the actual data manifold without explicit access. Chapter 5 provides an extensive analysis on the usage of DEPs for synthesizing CFs along with elaborate comparisons over choices of priors, space of regularization,datasets and baseline methods.

3.7 Summary

In this chapter, we showed that DEPs are effective estimators of total uncertainties or generalization error in DNNs, and more importantly, can regularize model training

when trained jointly with the classifier. Interestingly, models obtained using this approach demonstrate strong generalization characteristics, OOD rejection capabilities. We also provided a brief introduction to the key applications of DEPs, their key properties that can enable model introspection through feature importances and counterfactual explanations.

Chapter 4

FEATURE IMPORTANCE ESTIMATION USING DIRECT ERROR PREDICTORS

In this chapter, we focus on the problem of analyzing deep neural networks (DNNs) by estimating the importance of their features. This can provide information on the relative importance of each feature in making a particular prediction. We identify several limitations in current methodologies for feature importance estimation, including computational complexity, high uncertainty, and inability to handle changes in real-world data distributions. To address these issues, we propose a new approach that produces more robust and accurate feature importance estimates while also being computationally efficient. We show that DEP introduced in chapter 3, can be adapted for estimating feature importances. We revisit the training paradigm and elaborate upon the causal objectives adopted to explain a prediction using DEP. We provide results from a series of experiments and benchmarks. Our findings include key insights on the effectiveness of DEP for this purpose

4.1 Problem Setup

With the increased adoption of machine learning (ML) models in critical decision-making, post-hoc interpretability techniques are often required to enable decision-makers understand and trust these models. The *black-box* nature of ML models in most real-world settings (either due to their high complexity or proprietary nature) makes it challenging to interrogate their functioning. Consequently, attribution methods,

which estimate the influence of different input features on the model output, are commonly utilized to explain decisions of such black-box models. Existing approaches for attribution, or more popularly feature importance estimation, range from sensitivity analysis (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), studying change in model confidences through input feature masking (Schwab and Karlen 2019) to constructing simpler explanation models (e.g. linear, tree- or rule-based) that mimic a black-box model (Schwab and Hlavacs 2015; Lakkaraju et al. 2019).

Though sensitivity analysis techniques such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) are routinely used to explain individual predictions of any black-box classifier, they are computationally expensive. This challenge is typically handled in practice by constructing a *global* set of explanations using a sub-modular pick procedure (Ribeiro, Singh, and Guestrin 2016). On the other hand, despite being scalable, methods that construct simpler explanation models (Lakkaraju et al. 2019) are not guaranteed to match the behavior of the original model. While the recently proposed CXPlain (Schwab and Karlen 2019) addresses the scalability issue of feature masking methods, they are specific to the type of masking (e.g., zero masking) and the explainer needs to be re-trained if that changes (e.g., mean masking). Finally, and most importantly, it has been well documented that current approaches are highly sensitive to distribution shifts (Lakkaraju, Arsov, and Bastani 2020) and vulnerable to even small perturbations. Recently, Lakkaraju *et al.* (Lakkaraju, Arsov, and Bastani 2020) formalized this problem for the case of model mimicking approaches, and showed how adversarial training can be used to produce consistent explanations. In CXPlain, Schwab *et al.* proposed an ensembling strategy to effectively augment explanations with uncertainty estimates to better

understand the explanation quality. However, they did not study the consistency of inferred explanations under distribution shifts.

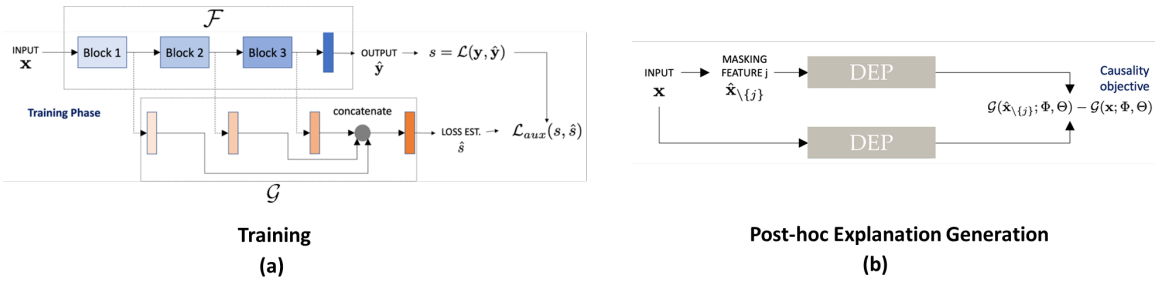
In this chapter, we propose PRoFILE (Producing Robust Feature Importances using Loss Estimates), a novel feature importance estimation method that is highly accurate, computationally efficient, consistent with the black-box model being explained and robust under distribution shifts. The key idea of our approach is to jointly train a DEP while building the predictive model using the objectives defined in Chapter 3, and generate post-hoc explanations by measuring the influence of input features on the model output using a causal objective defined on the loss estimates. Note that, once trained, the DEP can also be treated as a black-box. Interestingly, we find that the DEP is easier to train than obtaining calibrated uncertainty estimates, yet produces higher fidelity explanations. Unlike existing approaches, PRoFILE requires no re-training at explanation time and natively supports arbitrary masking strategies. Finally, using a variety of benchmarks, we show that the resulting explanations are robust under regimes of distribution shifts where the predicitive model generalizes to.

In summary, our contributions are:

- A computational efficient feature masking-based explainability method that is agnostic to the type of masking;
- Our approach is applicable to any data modality, deep architecture, or task;
- Experiments on a wide variety of both synthetic and real world data demonstrating the efficacy of PRoFILE under distribution shifts.

4.2 Related Work

Post-hoc explanation methods are the *modus-operandi* in interpreting the decisions of a black box model. Broadly, these approaches can be categorized as methods that generate explanations based on (a) sensitivity analysis; (b) gradients between the output and the input features; (c) change in model confidence through input feature masking; and (d) constructing simpler explanation models that can well approximate the black box predictor. LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) are two popular sensitivity analysis methods, and they produce sample-wise, local explanations based on regression models by measuring the sensitivity of the black-box to perturbations in the input features. However, these methods are known to involve significant computational overheads. On the other hand, Saliency Maps (Simonyan, Vedaldi, and Zisserman 2013), Integrated Gradients (Sundararajan, Taly, and Yan 2017), Grad-CAM (Selvaraju et al. 2017), DeepLIFT (Shrikumar, Greenside, and Kundaje 2017) and a gradient based version of SHAP - DeepSHAP (Lundberg and Lee 2017), are examples of gradient-based methods which are computationally effective. More recently, Schwab *et al.* proposed CXPlain (Schwab and Karlen 2019) and Attentive Mixture of Experts (Schwab, Miladinovic, and Karlen 2019), which are popular examples for methods that estimate model confidences through feature masking. Trained using a Granger causality-based objective (Granger 1969), these methods produce attention scores reflective of the feature importances, at a significantly lower computational cost. Finally, global explanation methods rely on mimicking the black-box using simpler explainer functions. For instance, ROPE (Lakkaraju, Arsov, and Bastani 2020) and MUSE (Lakkaraju et al. 2019) construct scalable, simple linear models and decision sets, to emulate



Training
(a)

Post-hoc Explanation Generation
(b)

Figure 14. An Illustration of Our Approach, PRoFILE, for Feature Importance Estimation. (Top) During the Training Phase, We Train a DEP along with the Predictive Model; (Bottom) We Use a Granger Causality-based Objective to Generate Post-hoc Explanations Using the Loss Estimates with No Re-training.

black-box models. An inherent challenge of this class of approaches is that the simple explainers are not guaranteed to match the behavior of the original model.

While these classes of methods vary in terms of their fidelity and complexity, a common limitation that has come to light recently is that explanations from most existing methods are associated with large uncertainties (Y. Zhang et al. 2019) and are not robust under distribution shifts. Recently, Lakkaraju *et al.* (Lakkaraju, Arsov, and Bastani 2020) explored the use of adversarial minmax training to ensure that the mimicking explainer model is consistent with the black-box under adversarial perturbations. In contrast, we find that, without any adversarial training, PRoFILE estimates feature importances robustly under distribution shifts, is computationally scalable compared to existing local explanation methods, and produces higher fidelity explanations.

4.3 Approach

We adopt an identical strategy as discussed in Chapter 3 to jointly train the predictor alongside the DEP (Figure 14 left). We utilize either the *contrastive* or *dropout calibration* objectives to train such models.

4.3.1 Feature Importance Estimation

Given the DEP $\mathcal{G}$ (Figure 14 right), we estimate the feature importance using a Granger causality-based objective, similar to (Schwab and Karlen 2019). The Humean definition of causality adopted by Granger (Granger 1969) postulates that a causal relationship exists between random variables x_j and y, i.e., $x_j \rightarrow y$, if we can better predict using all available information than the case where the variable x_j was excluded. This definition is directly applicable to our setting since it satisfies the key assumptions of Granger causality analysis, our data sample $\mathbf{x}$ contains all relevant variables required to predict the target and $\mathbf{x}$ temporally precedes $\mathbf{y}$. Mathematically,

$$\Delta\epsilon_{\mathbf{x},j} = \epsilon_{\mathbf{x}\setminus\{j\}} - \epsilon_{\mathbf{x}}, \tag{4.1}$$

where ϵ denotes the model error. For a sample $\mathbf{x}$, we can compute this objective for each feature j to construct the explanation. As showed in Figure 14(bottom), we use the DEP to measure the predictive model's error in the presence and absence of a variable x_j to check if x_j causes the predicted output.

There are a variety of strategies that can be adopted to construct $\mathbf{x} \setminus \{j\}$. In the simplest form, we can mask the chosen feature by replacing it with zero or a pre-specified constant. However, in practice, one can also adopt more sophisticated

masking strategies that take into account the underlying data distribution (Janzing et al. 2013; Štrumbelj, Kononenko, and Šikonja 2009). Interestingly, our approach is agnostic to the masking strategy and the DEP can be used to compute the causal objective in Eqn.(4.1) for any type of masking. In contrast, existing approaches such as CXPlain requires re-training of the explanation model for the new masking strategy.

Since DEP is jointly trained with the main predictor, our approach does not require any additional adversarial training as done in (Lakkaraju, Arsov, and Bastani 2020) to ensure that the explanations are consistent with the black-box. It must be noted that adversarial training for improving the robustness of the black box is independent of the design of PRoFILE. As the black box becomes more robust to arbitrary shifts/adversarial perturbations, we expect PRoFILE explanations to still be consistent. Existing works in the active learning literature have also found that the loss function (Yoo and Kweon 2019) or its gradients (Ash et al. 2020) effectively capture the inherent uncertainties in a model and hence can be used for selecting informative samples. Using a similar argument, we show that even though our causal objective is similar to CXPlain, our approach more effectively generalizes to even complex distribution shifts where CXPlain fails.

4.4 Experimental Setup

4.4.1 Datasets

We consider a suite of synthetic and real-world datasets to evaluate the fidelity and robustness of our approach. For the fidelity comparison study under standard testing conditions, we use the: (a) UCI Handwritten Digits dataset, (b) OpenML

benchmarks (Vanschoren et al. 2013), Kropt, Letter Image Recognition, Pokerhand and RBF datasets and (c) CIFAR10 image classification dataset (Krizhevsky and Hinton 2009). The dimensionality of the input data ranges from 10 to 64 and the total number of examples varies between 1797 and 13750 for different benchmarks. For each of the UCI and OpenML datasets, we utilized $\sim 90\%$ of the data while for CIFAR10, we used the prescribed dataset of 50K RGB images of size 32×32 for training our model. For the robustness study, we used the following datasets: (a) *Synthetic dataset*: In order to study the impact of distribution shifts on explanation fidelity, we constructed synthetic data based on correlation and variance shifts to the data generation process defined using a multi-variate normal distribution. More specifically, we generated multiple synthetic datasets of 5K samples, where the number of covariates was randomly varied between 10 and 50. In each case, the samples were drawn from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mu_{ii} = \alpha, \Sigma_{ii} = 1$ and $\Sigma_{ij} = \beta$ and the values α, β (the correlation between any two variables) were randomly chosen from the uniform intervals $[-2, 2]$ and $[-1, 1]$ respectively. The label for each sample was generated using their corresponding quantiles (i.e., defining classes separated by nested concentric multi-dimensional spheres). To generate correlation shifts, we created new datasets following the same procedure, but using a different correlation $\bar{\beta} = \beta + \delta_\beta$. Here, δ_β was randomly drawn from the uniform interval $[-0.2, 0.2]$. Next, we created a third dataset to emulate variance shifts, wherein we changed the variance $\Sigma_{ii} = \Sigma_{ii} + \kappa$ and κ was drawn from the uniform interval $[0.25, 0.75]$. While the predictive model was trained only using the original dataset, the explanations were evaluated using both correlation- and variance-shifted datasets. We generated 10 different realizations with this process and report the explanation fidelity metrics averaged across the 10 trials; (b) *CIFAR10 to CIFAR10-C* (Hendrycks and Dietterich 2019): This is a popular

benchmark for distribution-shift studies, wherein we train the predictive model and DEP using the standard CIFAR10 dataset and generate explanations for images from the CIFAR10-C dataset containing wide-variety of natural image corruptions; and (c) *MNIST-USPS*: In this case, we train the predictive model using only the MNIST handwritten digits dataset (LeCun, Cortes, and Burges 2010) and evaluate the explanations on the USPS dataset (Hull 1994) at test time.

4.4.2 Baselines

We compared PROFiLE against the following baseline methods that are commonly adopted to produce sample-level explanations. All baseline methods considered belong to the class of post-hoc explanation strategies which aim to construct interpretable models that can approximate the functionality of any black-box predictor.

(a) LIME (Ribeiro, Singh, and Guestrin 2016): LIME constructs linear models, which can locally approximate a black box predictor, by fitting a weighted regression model around the sample to be explained based on variants of the sample obtained by perturbing or zero-masking the input features. The intuition is that the post-hoc regression model obtained is reflective of the sensitivity of the black-box predictor to the modifications in the input features. The coefficients of the obtained post-hoc model serve as attribution scores for each feature in the given sample.

(b)SHAP (Lundberg and Lee 2017): SHAP determines the feature attribution scores for a sample by marginalizing the individual contributions of every feature towards a prediction. SHAP, more specifically KernelSHAP, fits a local regression model around the sample to be explained using multiple realizations of the sample by zero masking single or groups of features. A fundamental difference between LIME and SHAP lies

in the SHAP kernel used, which is a function of the cardinality of the features present in a group. The coefficients of the obtained model are the SHAPLey attribution scores for every feature in the given sample.

(c) CXPlain (Schwab and Karlen 2019): This determines feature attribution scores by training a post-hoc model that learns to approximate the distribution of Granger causal errors (Granger 1969), i.e., the difference between the black-box prediction loss when no feature is masked and the loss when features are zero-masked one at a time. The feature attribution scores obtained from the model are thus reflective of the global distribution of the causality based error metric. Similar to (Schwab and Karlen 2019), we use an MLP and a U-Net model as the post-hoc explainer for the non-image and the image datasets respectively in our experiments.

(d) Deep Shap (Lundberg and Lee 2017) DeepSHAP is a fast and scalable approximation of SHAP and also closely related to the DeepLIFT algorithm. We utilize this baseline on datasets where LIME and SHAP were expensive to run.

4.4.3 Evaluation Metric

To evaluate the explanation fidelity, we utilize the commonly used difference in log-odds metric, which is a measure of change in prediction when $k\%$ of the most relevant features in the input data are masked.

$$\Delta\text{log-odds} = \text{log-odds}(\text{p}_{\text{ref}}) - \text{log-odds}(\text{p}_{\text{masked}}) \tag{4.2}$$

Here log-odds(p) = $\log(\frac{\text{p}}{1-\text{p}})$ and p_{ref} is the reference prediction probability of the original data and p_{masked} refers to the prediction probability when a subset of features are masked. A higher value for Δlog-odds implies higher fidelity of the feature importance estimation. More specifically, for: (a) *Non-Image Datasets.* We sort

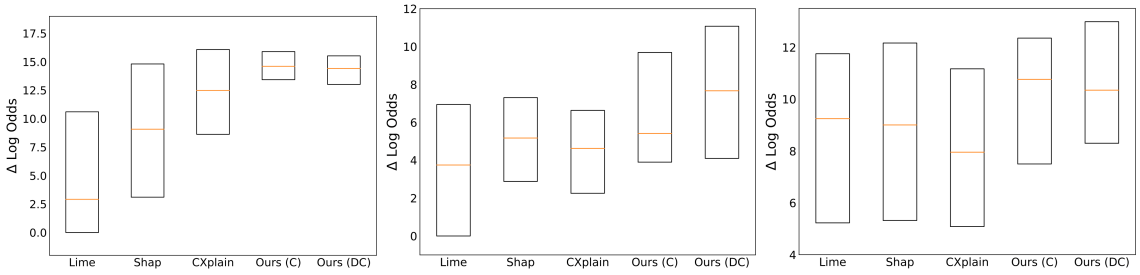
the feature attribution scores obtained from the explainability method (PROFiLE and baselines) and zero mask the top $k\%$ important features in the input sample to evaluate the metric, and (b) *Image Datasets*. We use the SLIC (Achanta et al. 2012) segmentation algorithm to generate super-pixels, which are then used to compute the feature importance scores. For CXPlain and DeepSHAP, we aggregate the pixel-level feature importance scores to estimate attributions for each super-pixel.

4.4.4 Hyper-parameters

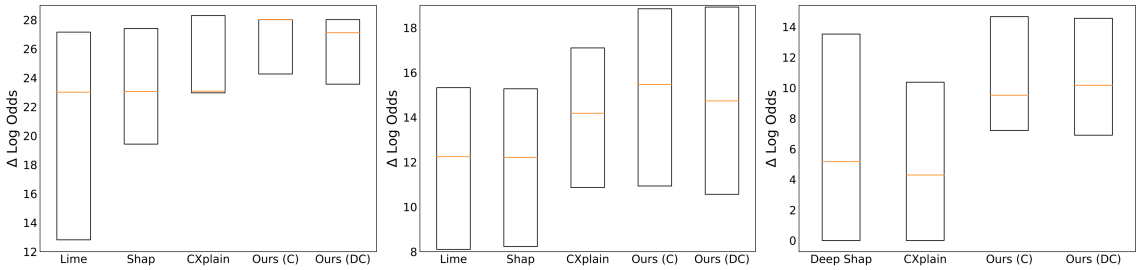
For all non-imaging datasets, the black-box model was a 5 layer MLP with ReLU activations, each fully-connected (FC) layer in DEP contained 16 units. In the case of CIFAR-10, we used the standard ResNet-18 architecture, and DEP used outputs from each residual blocks (with fully connected layers containing 128 hidden units). Finally, for the MNIST-USPS experiment, we used a $3-$layer CNN with 2 FC layers. DEP was designed to access outputs from the first 4 layers of the network and utilized FC layers with 16 units each. All networks were trained using the ADAM optimizer with a learning rate 0.001 and batch size 128.

4.5 Results and Findings

In this section, we present empirical studies to compare PROFiLE against popular baselines using both non-image and image benchmark datasets. More importantly, we evaluate the fidelity of the inferred explanations under challenging distribution shifts and demonstrate the effectiveness of PROFiLE.



(a) UCI Digits ($k = 15\%$) (b) Kropt ($k = 15\%$) (c) RBF ($k = 15\%$)



(d) Poker ($k = 15\%$) (e) Handwritten Letters ($k = 15\%$) (f) CIFAR-10 ($k = 15\%$)

Figure 15. Comparing the Fidelity of Feature Importances Inferred Using Different Methods. We Use the δlog-odds Score (Higher the Better) Obtained by Masking the Most Influential Input Features. For Each of the Datasets, the Ratio of Features Masked Is Also Included in Parentheses. Across All Benchmarks, PRoFILE is Consistently Superior over the Baselines.

4.5.1 Impact of DEPs on Fidelity of Feature Explanations

Figure 15 illustrates the Δlog-odds obtained using PRoFILE with both the learning strategies (Ours(C) and Ours(DC)) in comparison to the baselines. Note that, for the UCI and OpenML datasets, we used the held-out test set for our evaluation (90-10 split), while for CIFAR-10, we used 50 randomly chosen test images for computing the fidelity metric. While PROFiLE and CXPlain are scalable to larger test sizes, the small subset of test samples was used to tractably run the other baselines. For each dataset, we show the median (orange), along with the 25^{th} and the 75^{th} percentiles, of the Δlog-odds scores across the test samples. We find that PRoFILE consistently outperforms the

existing baselines on all benchmarks. In particular, both contrastive training and dropout calibration strategies are effective and perform similarly in all cases. The improved fidelity can be attributed directly to the efficacy of DEP and the causal objective used for inferring the feature attribution. In comparison, both LIME and SHAP produce lower fidelity explanations, while also being computationally inefficient. Interestingly, though CXPlain also uses a causal objective similar to us, the resulting explanations are of significantly lower fidelity. In terms of computational complexity for generating post-hoc explanations, PROFiLE which requires p evaluations (number of features that need to be masked and can be parallelized) of DEP and is only marginally more expensive than CXPlain.

4.5.2 Impact on DEP Feature Explanations under Distribution Shifts

Following our observations on the behavior of DEP, we now evaluate the fidelity of PRoFILE explanations in those scenarios. Figure 16 illustrates the median Δlog-odds and error bars obtained by masking the top 25% of features on 10 realizations of the synthetic dataset. In particular, we show the results for the held-out correlation and variance shifted data, while the models were trained only using the original synthetic data. We find that by utilizing a pre-trained DEP, PROFiLE significantly outperforms the baselines, even under complex shifts, indicating the robustness of our approach. Similar to the findings in (Lakkaraju, Arsov, and Bastani 2020), we note that the widely-adopted baselines are not immune to shifts. Figure 17 shows a detailed comparison of Δlog-odds for the CIFAR10-C dataset. Note, we show the median, 25^{th} and 75^{th} percentiles. We find that PRoFILE consistently achieves superior fidelity,

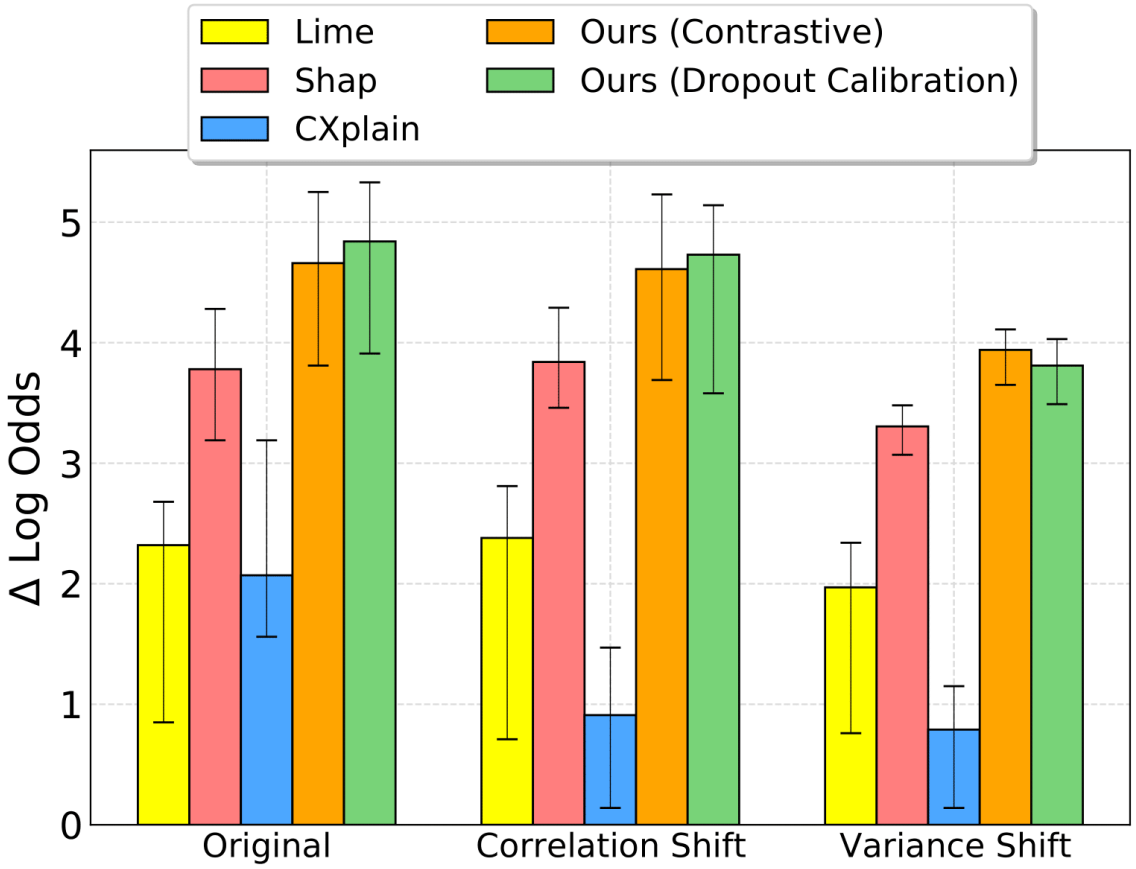


Figure 16. Using a Synthetic Dataset to Study the Robustness of Explanations Obtained Using Different Approaches, under Correlation and Variance Shifts. We Mask the Top 25% of Features in the Data to Obtain the δlog-odds Scores.

when compared to existing baselines, except in the case of *glass blur* where the scores are comparable.

Figure 18 shows examples of explanations obtained using PRoFILE on the USPS and CIFAR10-C datasets. We observe from Figure 18 (top) that our method adapts well across domains to identify critical pixels that characterize class-specific decision regions. Interestingly, these are examples where digit 8 is suitably masked by PRoFILE (only 5% of pixels) to be predicted as one of the other classes sharing the decision boundary. It can also be seen from Figure 18 (bottom) that PRoFILE explanations

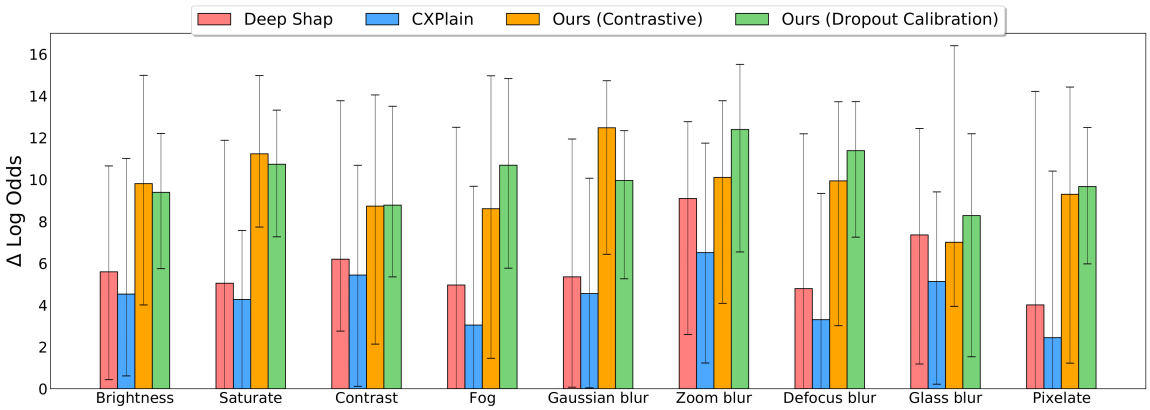


Figure 17. CIFAR-10C Dataset:- We Study the Fidelity of Explanations Generated on Different Types of Corrupted Images Using DEP Trained on the Original CIFAR-10 Data.

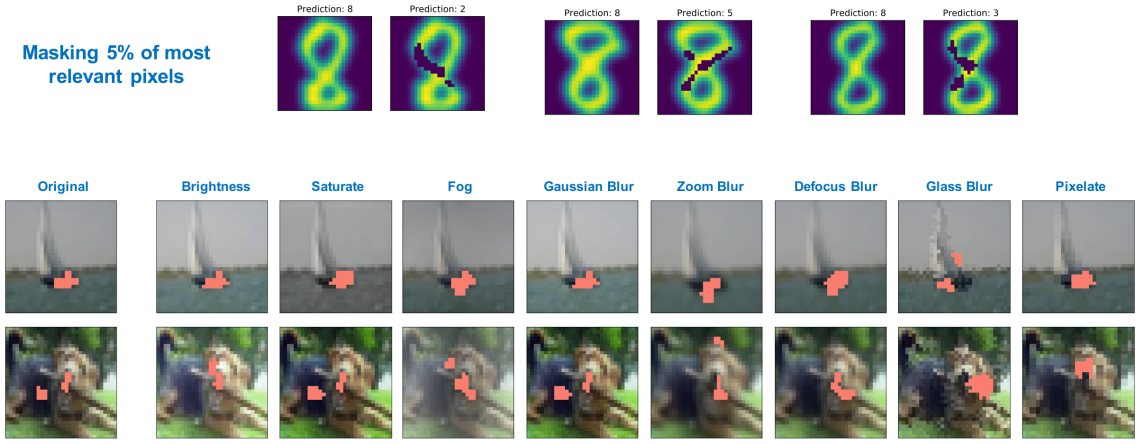


Figure 18. Examples of Explanations Generated Using PROFiLE (with Dropout Calibration) on USPS and CIFAR-10C Datasets Using Models Trained with MNIST and CIFAR-10 Respectively.

obtained under different domain shifts are consistent. In all cases except *glass blur*, it identifies the hull of the boat and the mouth of the dog as critical features. These observations strongly corroborate with the performance improvements in Figure 17.

4.6 Summary

In this chapter, we developed PROFiLE, a novel post-hoc feature importance estimation method based on DEPs applicable to any data modality or architecture. Using the pre-trained DEP along with a causality based objective, PROFiLE can accurately estimate feature importance scores that are immune to a wide variety of distribution shifts. Through extensive experimental studies on different data modalities, we demonstrated that PROFiLE provides higher fidelity explanations, is robust under real-world distribution shifts and is computational effective when compared to commonly adopted feature importance estimation methods.

Chapter 5

DESIGNING COUNTERFACTUAL GENERATORS USING DEEP MODEL INVERSION

In this chapter, we consider the task of synthesizing meaningful counterfactual (CF) explanations to introspect DNN models. We focus on a challenging scenario where we have access only to the trained deep classifier and not the actual training data even in the form generative models. The task of synthesizing CFs can be formulated as an inverse problem with a goal to introduce discernible changes in an image consistent with the target hypothesis. While the problem of inverting deep models to synthesize images from the training distribution has been explored, we identify the limitations of such methods to be directly adopted for CF explanations. We close this gap by proposing new strategies from first principles for training on the fly, counterfactual generators that produces highly interpretable CFs. Specifically, we analyze the role of DEP introduced in Chapter 3 in controlling the inversion by producing CFs that are evidenced by the data regimes used for training the classifier. We systematically report our experiments and benchmarks and provide key observations.

5.1 Problem Setup

With the growing need for deploying deep black-box models into critical decision-making, there is an increased emphasis on explainability methods that can reveal intricate relationships between data signatures (e.g., image features) and predictions. In this context, the so-called counterfactual explanations (Verma, Dickerson, and

Hines 2020) that synthesize small, interpretable changes to a given image while producing desired changes in model predictions to support user-specified hypotheses (e.g., progressive change in predictions) have become popular. Though counterfactual explanations provide more flexibility over conventional techniques, such as feature importance estimation (Selvaraju et al. 2017; Lakkaraju, Arsov, and Bastani 2020; Shrikumar, Greenside, and Kundaje 2017; Ribeiro, Singh, and Guestrin 2016; Ribeiro, Singh, and Guestrin 2018), by exploring the vicinity of a query image, an important requirement to produce meaningful counterfactuals is to produce discernible local perturbations (for easy interpretability) while being realistic (close to the underlying data manifold). Consequently, existing approaches rely extensively on pre-trained generative models to synthesize plausible counterfactuals (Verma, Dickerson, and Hines 2020; Looveren and Klaise 2021; Dhurandhar et al. 2018; Sumedha et al. 2020; Goyal et al. 2019). By design, this ultimately restricts their utility to scenarios where one cannot access the training data or pre-trained generative models, for example, due to privacy requirements commonly encountered in many practical applications.

In this chapter, we focus on the problem where we have access only to trained deep classifiers and not the actual training data or generative models. Synthesizing images from the underlying data distribution by inverting a deep model, while not requiring access to training data, is a well investigated topic of research.

For e.g., Deep Dream (Mordvintsev, Olah, and Tyka 2017) synthesizes class-conditioned images by manipulating a noisy image directly in the space of pixels (or more formally *Image Space Optimization* (ISO)) constrained by image priors such as total variation (Mahendran and Vedaldi 2016) to regularize this ill-posed inversion. However, Deep Dream is known to produce images that look unrealistic, often very different from the training images, thus limiting their use in practice. Consequently,

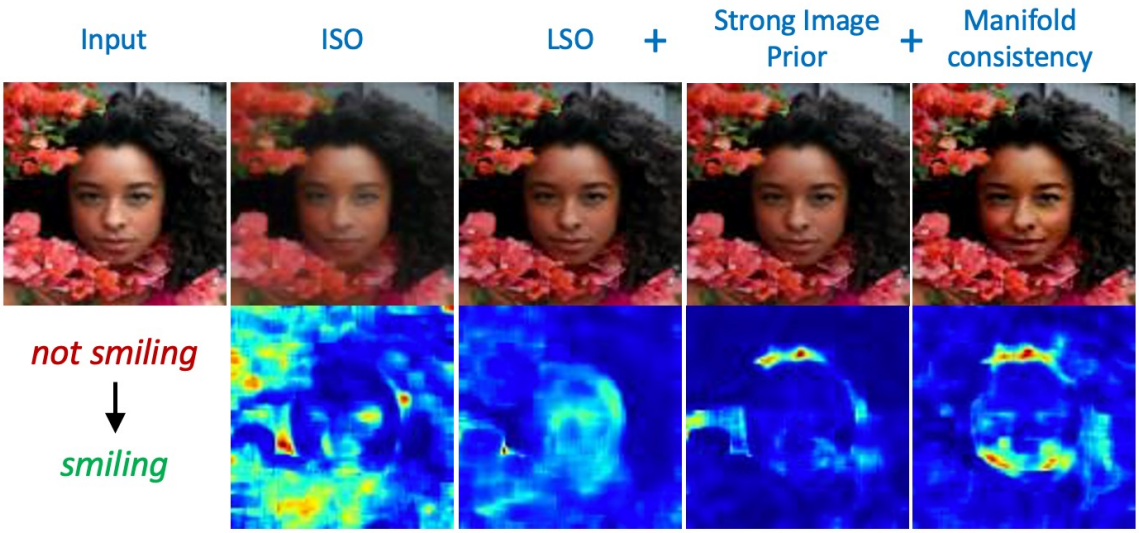


Figure 19. We Propose DISC, a Deep Model Inversion Approach for Query-based CF Generation. Using a Strong Image Prior (INR in This Example) and Our Manifold Consistency Constraint, along with a Progressive Optimization Strategy, DISC introduces Discernible yet Semantically Meaningful Changes (Rightmost) to the Query Image.

Yin *et al.*, proposed *DeepInversion* (Yin et al. 2020) that performs image synthesis in the latent space of a pre-trained classifier (*Latent Space Optimization* (LSO)) and leverages layer-specific statistics (from batchnorm (Ioffe and Szegedy 2015)) to constrain the images to be consistent with the training data distribution. This was showed to produce higher-quality images, particularly in the context of performing knowledge distillation (Geoffrey, V., and D. 2015) using the synthesized images.

Our Work: In contrast, this work aims to develop a deep model inversion approach that generates counterfactual explanations by exploring the vicinity of a given query image, instead of synthesizing an arbitrary realization from the entire image distribution. As illustrated in the example in Figure 19, existing deep inversion methods are ineffective when natively adopted for counterfactual generation. Due to use of weak priors, and the severely ill-posed nature of the problem, it introduces irrelevant

pixel manipulations that easily satisfy the desired change in prediction. Hence, we propose DISC (Deep Inversion for Synthesizing Counterfactuals) that improves upon conventional deep model inversion by utilizing: (i) stronger image priors through the use of deep image priors (Ulyanov, Vedaldi, and Lempitsky 2018) (DIP) and implicit neural representations (Sitzmann et al. 2020) (INR); (ii) a novel *manifold consistency* objective enforced using DEP that ensures the counterfactual remains close to the underlying manifold; and (iii) a progressive optimization strategy to effectively introduce discernible, yet meaningful, changes to the query image.

From Figure 19, we find that our approach produces meaningful image manipulations, in order to change the prediction to the *smiling* class, while other deep inversion strategies cannot. Using empirical studies, we show that DISC consistently produces visually meaningful explanations, and that the counterfactuals from DISC are effective at learning model decision boundaries and are robust to unknown test-time corruptions.

Our Contributions

1. A general framework to produce counterfactuals on-the-fly using deep model inversion;
2. Novel objectives to ensure consistency to the data manifold. We explore two different strategies based on direct error prediction (Yoo and Kweon 2019; JJ. Thiagarajan et al. 2021) and deterministic uncertainty estimation (Van Amersfoort et al. 2020);
3. A progressive optimization strategy to introduce discernible changes to a given query image, while satisfying the manifold consistency requirement;
4. A *classifier discrepancy* metric to evaluate the quality of counterfactuals;

5. Empirical studies using natural image and medical image classifiers to demonstrate the effectiveness of DISC over a variety of baselines and ablations.

5.2 Related Work

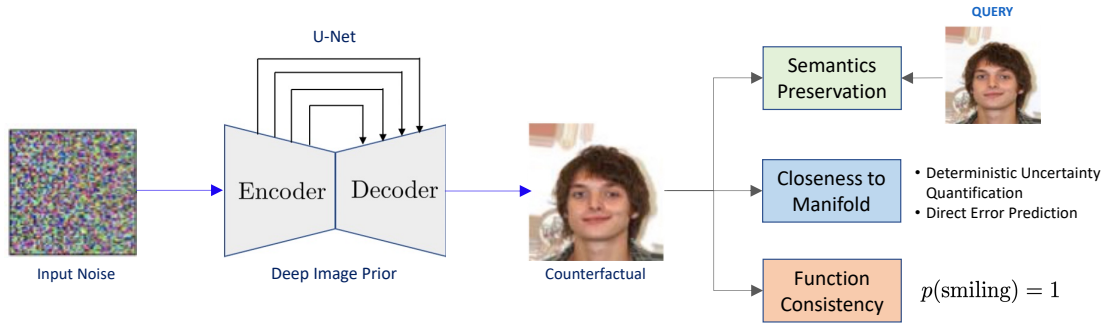
(a) Image Synthesis from Classifiers: Inverting a pre-trained deep model is a popular strategy for image synthesis in scenarios where there is no access to the underlying training data. While deep model inversion-based methods such as *Deep Dream* (Mordvintsev, Olah, and Tyka 2017) and *DeepInversion* (Yin et al. 2020) have been successful in generating class-conditioned images, there have been other extensions to such approaches. For example, Dosovitskiy *et al.* (Dosovitskiy and Brox 2016) proposed to invert representations of a pre-trained CNN to obtain insights about what a deep classifier has learned. On similar lines, Mahendran *et al.* (Mahendran and Vedaldi 2016; Mahendran and Vedaldi 2015) addressed the problem of pre-image recovery, which in essence attempts to recover an arbitrarily encoded representation (in the latent space of a classifier) to a realization on the (unknown) image manifold and to enable model diagnosis. Ulyanov *et al.* (Ulyanov, Vedaldi, and Lempitsky 2018) improved upon this ill-posed inversion by utilizing a strong image prior in the form of *deep image priors* (DIP). They also explored the related problem of activation-maximization (Ulyanov, Vedaldi, and Lempitsky 2018), where the goal is to generate an image that maximizes the activation of a given output neuron in a pre-trained classifier, and demonstrated the effectiveness of DIP. Despite the effectiveness of the deep model inversion methods for image synthesis, we find that such methods cannot be natively adopted for CF generation and are insufficient for producing meaningful pixel manipulations to a given query image.

(b) Counterfactual Generation: Existing methods extensively rely on generative models to provide counterfactuals that explain the decisions of a black-box model. Examples including CounteRGAN (Nemirovsky et al. 2020), Counterfactual Generative Networks (CGNs) (Sauer and Geiger 2021) and the methods reported in (Looveren and Klaise 2021; Dhurandhar et al. 2018), have clearly demonstrated the use of generative models in synthesizing high-quality CFs for any user-specified hypothesis on the predictions. However, this requirement of access to training data or generative models can be infeasible in practical scenarios, for e.g., restrictions arising due to privacy requirements. In contrast, our approach formulates the problem of counterfactual generation using deep model inversion, and can produce meaningful counterfactuals using only the trained classifier.

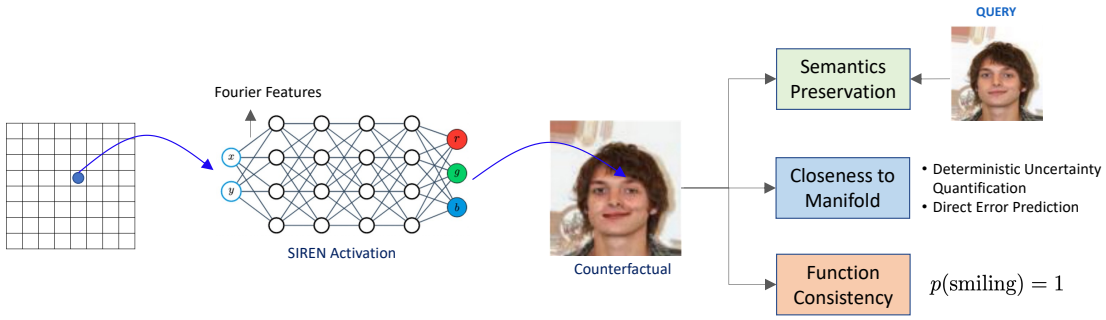
5.3 Approach

In this section, we describe our approach for counterfactual generation that improves upon deep model inversion, and introduce a new classifier discrepancy metric for evaluating CFs. There are four key components that are critical to designing classifier-based counterfactual generators: (i) choice of metric for *semantics preservation*; (ii) choice of image priors to regularize image synthesis; (iii) *manifold consistency* to ensure that the synthesized counterfactual lies close to the true data manifold; and (iv) progressive optimization strategy to introduce gradual meaningful changes to a query image. Figures 20a and 20b provide overview of DISC implemented with deep image prior and INR-based for regularizing the counterfactual optimization process.

In its simplest form, for a given query $\mathbf{x}$, a counterfactual explanation can be



(a) Deep Image Prior



(b) Implicit Neural Representations

Figure 20. Overview of Our Approach. For a Given Query, DISC trains an Image Generator, Which Can Be Implemented Using a Deep Image Prior (DIP) or Coordinate-based Neural Representations (INR), Based on Three Key Objectives: (i) Semantics Preservation; (ii) Manifold Consistency; And (iii) Function Consistency.

obtained as follows:

$$\arg\min_{\bar{\mathbf{x}}} d(\bar{\mathbf{x}}, \mathbf{x}) \quad \text{s.t.} \quad \mathcal{F}(\bar{\mathbf{x}}) = \bar{\mathbf{y}},\ \bar{\mathbf{x}} \in \mathcal{M}(\mathbf{x}), \tag{5.1}$$

where $\bar{\mathbf{x}} = \mathcal{C}(\mathbf{x})$ is a counterfactual explanation for $\mathbf{x}$, $\mathcal{F}$ is a pre-trained classifier model, $\mathcal{M}$ denotes the data manifold and $\bar{\mathbf{y}} = \mathbf{y} + \delta$ is the desired change in the prediction. The metric $d(.,.)$ measures the discrepancy between the query image and the counterfactual (i.e., semantics preservation).

5.3.1 Choice of Metric for Semantics Preservation

We can measure the discrepancy between a query and its CF explanation, $d(\bar{\mathbf{x}}, \mathbf{x})$, in the pixel space or in the latent space of the classifier. We now describe the high-level formulation of deep model inversion-based CF generation.

5.3.1.1 Image Space Optimization (ISO)

ISO for counterfactual generation involves the ill-posed optimization of an input $\bar{\mathbf{x}}$ directly in the pixel space to generate an image semantically similar to the query $\mathbf{x}$, while being consistent with a user-hypothesis on the prediction, $\bar{\mathbf{y}}$. Strategies such as Deep Dream (Mordvintsev, Olah, and Tyka 2017) perform ISO to synthesize artistic variations of images. Mathematically,

$$\arg\min_{\bar{\mathbf{x}}} d(\bar{\mathbf{x}}, \mathbf{x}) + \mathcal{R}(\bar{\mathbf{x}}) \quad \text{s.t.} \quad \mathcal{F}(\bar{\mathbf{x}}) = \bar{\mathbf{y}}. \tag{5.2}$$

where $\mathcal{R}(\bar{\mathbf{x}})$ is a suitable image prior to regularize the optimization.

5.3.1.2 Latent Space Optimization (LSO)

LSO refers to the ill-posed problem of inverting an arbitrarily encoded representation from the latent space of a deep classifier to a realization on the (unknown) image manifold. Let $\Psi_l(.)$ denote the l^{th} differentiable layer of the deep classifier. Then, counterfactual generation using LSO can be mathematically formulated as

$$\arg\min_{\bar{\mathbf{x}}} \sum_{l} d(\Psi_l(\bar{\mathbf{x}}), \Psi_l(\mathbf{x})) + \mathcal{R}(\bar{\mathbf{x}}) \quad \text{s.t.} \quad \mathcal{F}(\bar{\mathbf{x}}) = \bar{\mathbf{y}}. \tag{5.3}$$

Approaches such as DeepInversion (Yin et al. 2020) perform LSO for conditional image synthesis (though with distribution-level comparison instead of our sample-level comparison) and achieve visually superior images when compared to ISO approaches. Equation (5.3) reduces to (5.2) when $l = 0$ which corresponds to the input layer of the classifier.

5.3.2 Choice of Image Priors

As observed from (5.2) and (5.3), the choice of the regularizer or image prior $\mathcal{R}(.)$ is central towards regularizing and tractably solve this challenging inverse problem. A variety of image priors have been proposed in the literature, and we investigate the following in this work.

5.3.2.1 Total Variation + l_2

Total variation (Mahendran and Vedaldi 2015) (TV) is a popular regularizer that encourages images to contain piece-wise constant patches while the ℓ_2 norm regularizes the range and energy of the image to remain within a given interval. The TV norm and the ℓ_2 norm are given by:

$$\mathcal{R}_{TV}(\bar{\mathbf{x}}) = \sum_{i,j} \sqrt{(\bar{\mathbf{x}}_{i,j+1} - \bar{\mathbf{x}}_{i,j})^2 + (\bar{\mathbf{x}}_{i+1,j} - \bar{\mathbf{x}}_{i,j})^2}; \quad \mathcal{R}_{\ell_2}(\bar{\mathbf{x}}) = \sqrt{\sum_{i,j} \|\bar{\mathbf{x}}_{i,j}\|^2} \tag{5.4}$$

5.3.2.2 Deep Image Priors (DIP)

A *Deep Image Prior* (DIP) (Ulyanov, Vedaldi, and Lempitsky 2018) leverages the structure of an untrained, carefully tailored convolutional neural network (e.g.,

U-Net (Ronneberger, Fischer, and Brox 2015)) to generate images and solve a variety of ill-posed restoration tasks in computer vision. DIP has been found to produce high-quality reconstructions, based on the key insight that the structure of the network itself can act as a regularizer. Consequently, the synthesized image is re-parameterized in terms of the weights θ of the prior model f_θ i.e $\bar{\mathbf{x}} = f_\theta(\mathbf{z})$.

5.3.2.3 Implicit Neural Representations (INR)

We also considered an alternative approach based on INR, which provide a cheap and convenient way to learn a continuous mapping from the image coordinates to the pixel values (RGB). While they have been found to be effective for image/volume rendering and designing generative models (Chen, Liu, and Wang 2021), we explore their use in deep model inversion (details in appendix). We build upon two key results to design our INR-based counterfactual generators:

(i) Fourier mapping: Based on neural tangent kernel (NTK) theory, Tancik *et al.* (Tancik et al. 2020) showed that using Fourier mapping can recover high-frequency features in low-dimensional coordinate-based image reconstruction. Hence, we use a Fourier feature mapping z to featurize $2-$D input coordinates $\mathrm{v} \in [0,1]^2$ before passing them through a coordinate-based MLP:

$$z(\mathrm{v}) = [a_1 \cos\big(2\pi \mathrm{b}_1^T \mathrm{v}\big), a_1 \sin\big(2\pi \mathrm{b}_1^T \mathrm{v}\big), \cdots].$$

Using a set of randomly chosen sinusoids, this maps the input points to the surface of a high-dimensional hyper-sphere. Training the MLP network on these embedded points corresponds to kernel regression with the stationary composed NTK $h_{\mathrm{NTK}} \circ h_{\mathrm{z}}$, where h_{NTK} denotes the neural tangent kernel corresponding to the MLP;

(ii) SIREN activation: In (Sitzmann et al. 2020), Sitzmann *et al.* showed that

periodic activation functions are better suited for recovering natural images and their derivatives, when compared to standard activation functions. More specifically, SIREN uses a sinusoid activation $\Phi(\mathbf{x}) = \sin(\mathbf{W}\mathbf{x} + \mathrm{b})$. We find that using both a Fourier mapping coupled with SIREN activation leads to a very strong image prior.

5.3.3 Manifold Consistency

A key constraint in (5.1) that is not included in the formulations in (5.2), (5.3) is the manifold consistency, and interestingly, this is not inherently satisfied in deep model inversion. Consequently, even with a strong image prior, it can synthesize images that do not belong to the original data distribution. This can be particularly challenging when producing CFs that represent change in class labels, wherein one expects patterns specific to a target class to be emphasized. To address this challenge, we extend the formulation in (5.3) (and equivalently (5.2)) to include a manifold consistency constraint, that is defined directly based on the classifier, without assuming access to class-specific statistics in the latent space. More specifically,:

$$\arg\min_{\bar{\mathbf{x}}} \lambda_1 \sum_l d(\Psi_l(\bar{\mathbf{x}}), \Psi_l(\mathbf{x})) + \lambda_2 \mathcal{L}_{mc}(\bar{\mathbf{x}}; \mathcal{F}) + \lambda_3 \mathcal{L}_{fc}(\mathcal{F}(\bar{\mathbf{x}}), \bar{\mathbf{y}}). \tag{5.5}$$

The first term for *semantics preservation* is same as that of (5.3) and is used to ensure that the inherent semantics of the query image is retained in the generated counterfactual (implemented as the ℓ_2 error). The second term $\mathcal{L}_{mc}$ (*manifold consistency*) penalizes solutions that do not lie close to the data manifold and is designed by assuming access only to the classifier $\mathcal{F}$. The final loss term $\mathcal{L}_{fc}$ (*functional consistency*) ensures that the prediction for the counterfactual matches the desired target $\bar{\mathbf{y}}$, e.g., categorical cross entropy. As illustrated in Figure 21, the manifold consistency objective plays a central role in deep inversion-based CF generation. Even

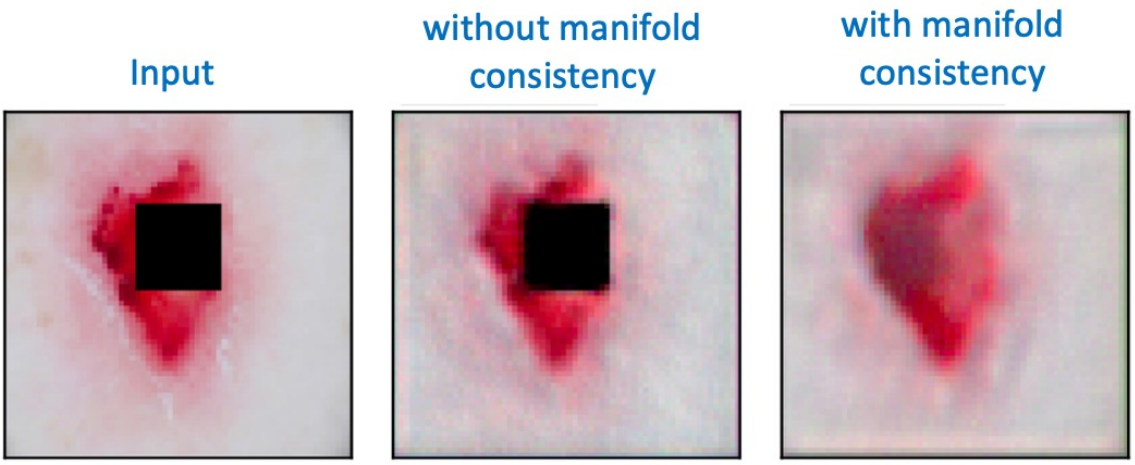


Figure 21. Need for Manifold Consistency. Without Explicitly Constraining the CFs to Lie Close to the True Manifold, Deep Inversion-based Generators Can Produce OOD Images (Missing Pixels) That Satisfy *Functional Consistency*. In Contrast, Our Approach Is Able to Create a More Faithful Explanation by Automatically Filling in Missing Pixels.

with a strong image prior (DIP in this example) and LSO, the generator produces out-of-distribution (OOD) CFs (with missing pixels) as guided by the semantic preservation term. Though the synthesized CF produces the desired class label with the classifier $\mathcal{F}$, the explanation is not interpretable. In contrast, including the $\mathcal{L}_{mc}$ term (using DEP explained next) leads to a meaningful counterfactual that automatically fills in the missing pixels. Note that, this example is different from prior-based *image inpainting* (Ulyanov, Vedaldi, and Lempitsky 2018), where a known mask is used to alter the loss function to recover the missing pixels. In this work, we explore two different strategies to implement $\mathcal{L}_{mc}$.

5.3.3.1 Direct Error Prediction (DEP)

In chapter 3, we found that a direct error predictor trained jointly with the classifier can be used to effectively detect distribution shifts and obtain accurate uncertainty estimates for a given sample. The DEP $\mathcal{G}$ (Figure 14) estimates the total generalization error a proxy of model confidence consistent with the underlying classifier. For all experiments provided in this chapter, we utilize the contrastive training objective defined in Chapter 3 to jointly train the predictor and DEP. Here, we view the DEP as a self-calibrated hypothesis tester and operate it in such a manner to synthesize a CF with a low loss. Using such an approach, we implement the manifold consistency term $\mathcal{L}_{mc}$ using DEP since the losses indicate regimes where the model fails to make an accurate prediction.

$$\mathcal{L}_{mc} = \|\mathcal{G}(\bar{\mathbf{x}}) - s^*\|_1, \tag{5.6}$$

where s^* denotes the target loss to be achieved. In our experiments, we set s^* as the median error estimate from DEP on a held-out validation set.

5.3.3.2 Deterministic Uncertainty Quantification (DUQ)

The recently proposed DUQ (Van Amersfoort et al. 2020) method is based on Radial Basis Function (RBF) (LeCun et al. 1998) networks and has been showed to be highly effective at OOD detection. In its formulation, a model is comprised of a deep feature extractor $\mathcal{F}$, an exponential kernel function along with a set of prototypical feature vectors (or *centroids*) for each class. DUQ is trained by optimizing the kernel distance between the features from $\mathcal{F}$ and the class-specific centroids and using a moving average process to update the centroids. Once DUQ is trained, the

uncertainty can be measured as the distance between the model output and the closest centroid. Details regarding DUQ training can be found in the appendix. In this work, we implement $\mathcal{L}_{mc}$ using a margin-based loss that maximizes the kernel similarity of the synthesized CF with the centroid of the target class, relative to the source class. Denoting the kernel similarity for a CF $\bar{\mathbf{x}}$ with the centroid for class $\mathbf{y}$ as $K(\mathcal{F}(\bar{\mathbf{x}}), \phi(\mathbf{y}))$, where $\phi(\mathbf{y})$ corresponds the pre-computed centroid for class y, we define:

$$\mathcal{L}_{mc} = \max\left(K(\mathcal{F}(\bar{\mathbf{x}}), \phi(\mathbf{y})) - K(\mathcal{F}(\bar{\mathbf{x}}), \phi(\bar{\mathbf{y}})) + \tau, 0\right), \tag{5.7}$$

which indicates that kernel similarity w.r.t. the target class $\bar{\mathbf{y}}$ should be greater than that with the source class $\mathbf{y}$ at least by the margin τ (set to 0.5 in our experiments).

5.3.4 Progressive Optimization

CF generation is a highly under-constrained problem, that even with a strong image prior and the manifold consistency constraint, it can easily converge to trivial solutions, i.e., irrelevant image manipulations. For example, one might expect to introduce large discernible changes by reducing the penalty λ_1 for semantics preservation. However, given the large solution space (defined by the number of parameters in the DIP/INR generator f_θ), this often leads to unrealistic images. To circumvent this, we propose to adopt a progressive optimization strategy that gradually increases the number of layers in f_θ to be optimized and steadily relaxing the penalty λ_1 (by factor κ) to allow for larger, yet interpretable, changes. More specifically, denoting the number of layers in f_θ by L, in each iteration we train the parameters of the first i layers (i is incremented by 1 in the subsequent iteration) while keeping the parameters of the remaining $L - i$ layers at their initial state (details on how the layers are

chosen for DIP and INR based generators can be found in the appendix). A similar strategy has been shown to be effective for ill-posed restoration tasks using large-scale generative models such as Style-GAN (Daras et al. 2021; Karras et al. 2020). An outline of this progressive optimization process is provided in the appendix. We find that such a progressive optimization leads to significantly better quality solutions allowing meaningful traversal from one class to another.

5.3.5 Evaluating Quality of CF Explanations using Classifier Discrepancy

A desired property in query-based explainers is that the synthesized changes are both interpretable and representative of the *target* class (*e.g., smiling*). To systematically evaluate the latter property, we propose the following synthetic experiment: Given a binary classifier $\mathcal{F}$ and training images, i.e., X_0, belonging to Class 0, we use our CF generator to synthesize examples for Class 1, i.e., $\bar{X}_1$ (using Class 0 images as input), and finally build a secondary classifier $\mathcal{F}^c$ using $[X_0, \bar{X}_1]$. The quality of the counterfactuals can thus be measured using the gap between the performance of $\mathcal{F}$ and $\mathcal{F}^c$ on a common test set. We refer to this score as *classifier discrepancy* (CD).

5.4 Experiment Setup

5.4.1 Datasets

(i) *CelebA Faces* (Z. Liu et al. 2015): This dataset contains 202,599 images along with a wide-range of attributes. For our experiments, we consider 3 different attributes, namely *smiling*, *bald* and *young*. Note, we train a classifier for predicting each of

the attributes independently. We report the results for *bald* and *young* attributes in the appendix; (ii) *ISIC 2018 Skin Lesion Dataset* (Codella et al. 2019): This lesion diagnosis challenge dataset contains a total of $10,015$ dermoscopic lesion images from the HAM10000 database (Tschandl, Rosendahl, and Kittler 2018). Each image is associated with one out of 7 disease states: Melanoma (MEL), Melanocytic nevus (MN), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BK), Dermatofibroma (DF) and Vascular lesion (VASC). Note, in all cases, we used a stratified $90 - 10$ data split to train the classifiers.

5.4.2 Model Design and Hyper-Parameters

(a) Classifier Design: For all experiments, we resized the images to size 96×96 and used the standard ResNet-18 architecture (He et al. 2016) to train the classifier model with the Adam optimizer (Kingma and Ba 2015), batch size 128, learning rate $1e-4$ and momentum 0.9. For the DEP implementation (Section 5.3.3.1), we performed average pooling on feature maps from each of the residual blocks in ResNet-18, and applied a linear layer of 128 units with ReLU activation. The hyper-parameters in were set at $\beta_1 = 1.0$ and $\beta_2 = 0.5$. For the case of DUQ, we set both the length scale parameter and the gradient penalty to 0.5.

(b) Image Generator Design: For generator design, the deep image prior used the standard U-Net architecture and input noise images drawn from the uniform distribution $\mathcal{U}[-1, 1]$. For INR, we chose 256 random sinusoids with frequencies b_i drawn from a Gaussian distribution with mean 0 and variance 100 to compute the Fourier mapping for the input coordinates.

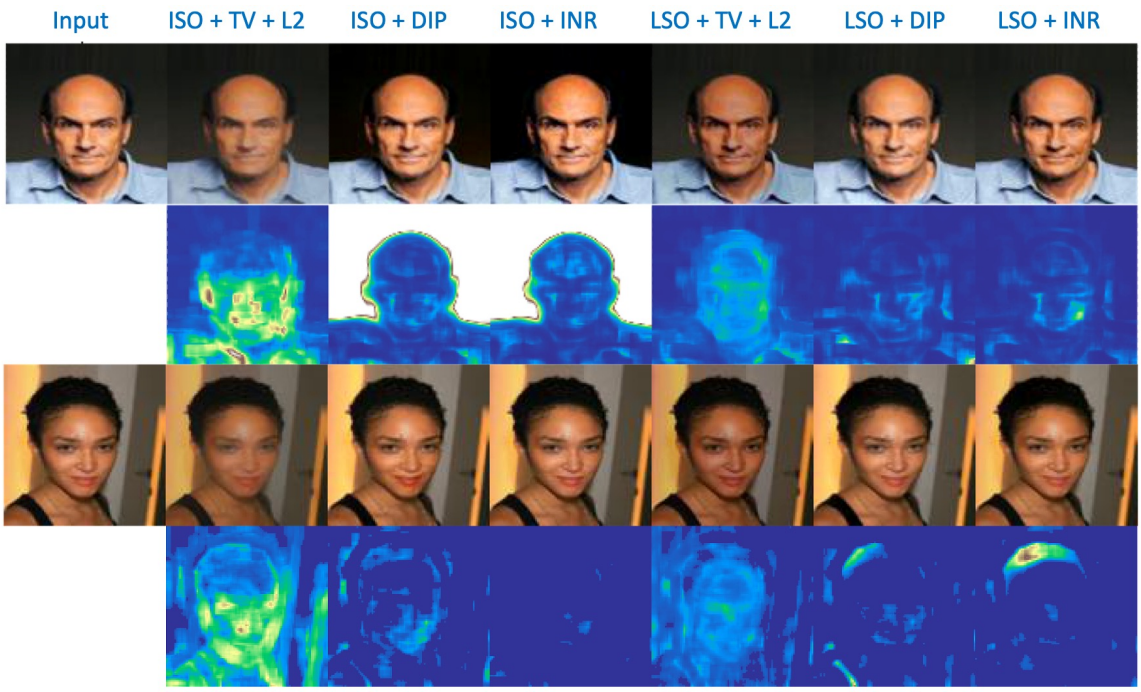


Figure 22. ISO vs LSO with Different Choices of Priors. Though None of the Image Priors Inherently Lead to Discernible Changes That Reflect the Properties of the Target *Smiling* Class, We Find That LSO with Strong Priors Produces Higher Quality Images Compared to ISO.

5.5 Results and Findings

5.5.1 Impact of Choosing Metrics for Semantics Preservation

An important design component in DISC is how we compute the semantic discrepancy between query $\mathbf{x}$ and CF $\bar{\mathbf{x}}$ - either in the pixel space using ISO or in the latent space of the classifier using LSO. We perform a comparative analysis of their behavior in CF generation using CelebA faces, in particular when manipulating an image from the *non-smiling* class to be classified as *smiling*, by varying the choice of image priors. As observed in Figure 22, though LSO produces higher quality images compared to ISO when using a weak image prior (TV + ℓ_2), that quality gap vanishes

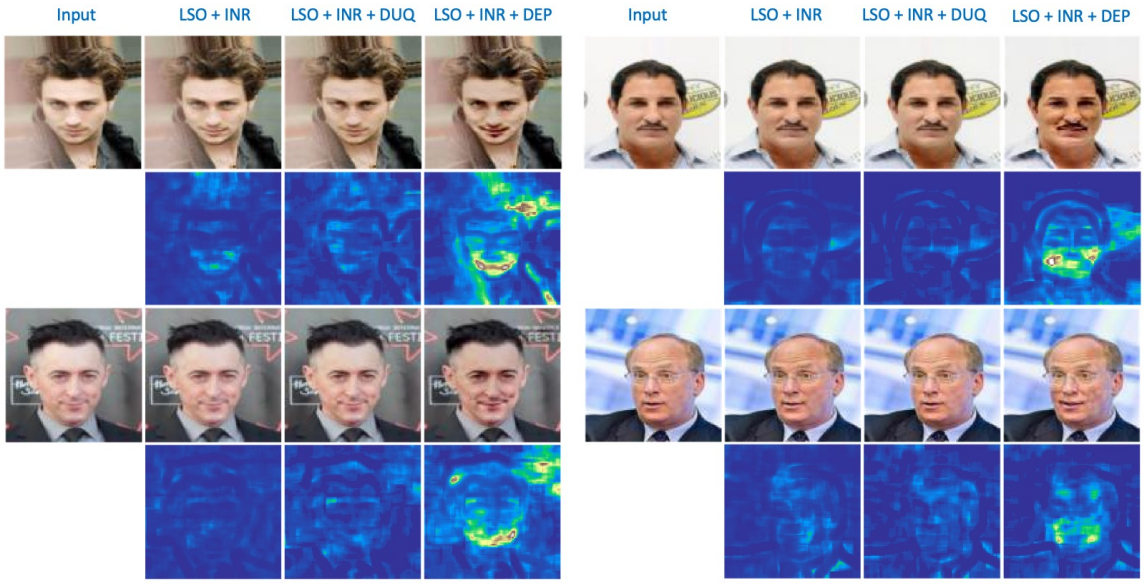


Figure 23. Importance of Manifold Consistency. The DEP Objective Significantly Improves over the Standard LSO (with No $\mathcal{L}_{mc}$) by Introducing Appropriate Pixel Manipulations near the Mouth and Cheeks in All Examples. In Contrast, We Find That DUQ-based Consistency Is Insufficient to Emphasize the Semantics of the *Smiling* Class as Seen in the Difference Images ($|\mathbf{x} - \bar{\mathbf{x}}|$).

with the use of a stronger prior, e.g., DIP. However, in terms of producing discernible changes that reflect the properties of the *smiling* class, neither approach is sufficient. In particular, ISO shows a higher risk of making minimal, irrelevant modifications (refer to difference images $|\mathbf{x} - \bar{\mathbf{x}}|$ in Figure 22) that drive the prediction to a desired label and hence, similar to (Yin et al. 2020), we recommend the use of LSO but with stronger image priors to design CF generators.

5.5.2 Importance of Manifold Consistency in Producing Meaningful Explanations

As showed in the previous experiment, deep model inversion does not produce discernible (and interpretable) image changes when applied for CF generation. In this context, we explore the impact of enforcing manifold consistency in DISC. In particular,

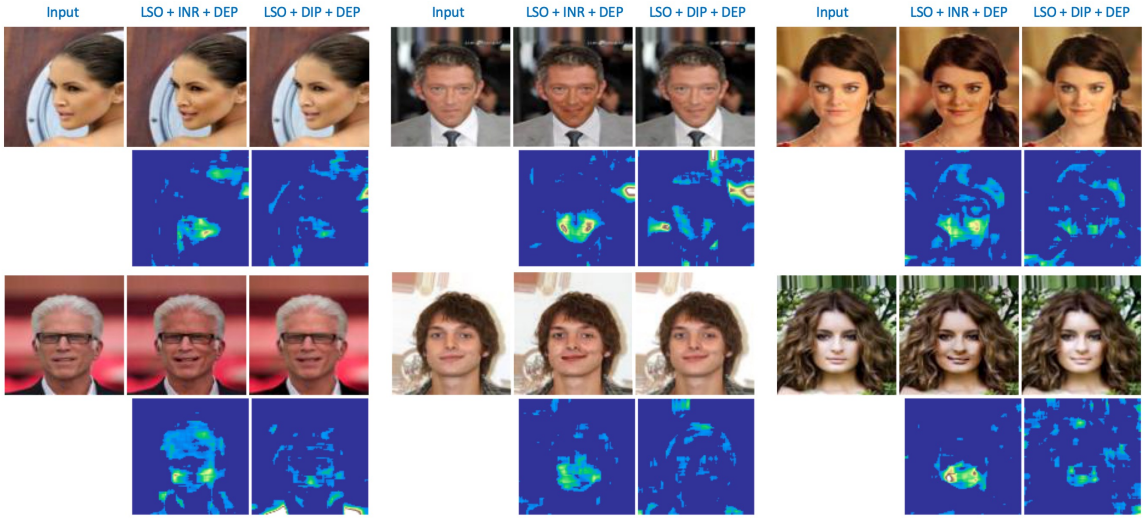


Figure 24. Comparison Between DIP and INR with DEP Manifold Consistency. Although Both DIP and INR Are Effective for LSO-based Model Inversion, We Find That INR Based Generators Produce Highly Concentrated and More Apparent Image Manipulations.

we compare the following LSO-based CF generation implementations (with INR prior): (i) no manifold consistency; (ii) DUQ-based $\mathcal{L}_{mc}$ from (5.7); and (iii) DEP-based $\mathcal{L}_{mc}$ from (5.6). From the results in Figure 23, we find that the DEP objective significantly improves over the standard LSO (with no $\mathcal{L}_{mc}$) by introducing appropriate pixel manipulations near the mouth and cheeks in all examples and clearly represents the underlying semantics of the *smiling* class. In contrast, the RBF network-based DUQ performs very similar to the *LSO + INR* baseline and this emphasizes the inability of the kernel similarity metric to detect mild distribution shifts in data, though it has been proven successful for detecting severely OOD samples.

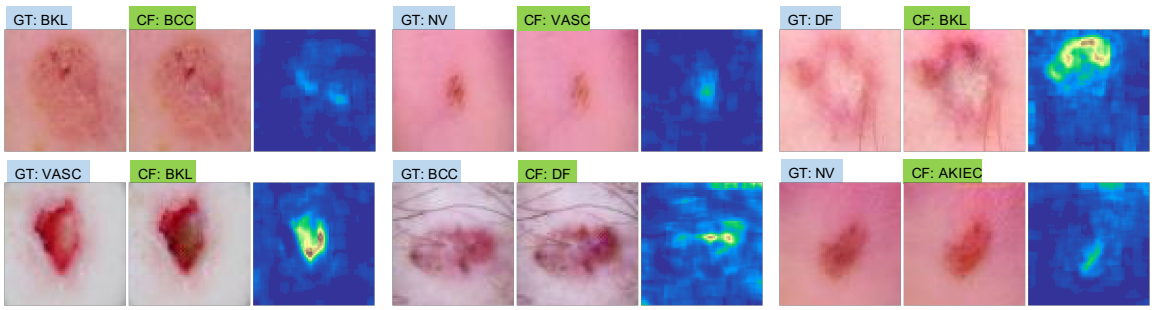


Figure 25. Observations on CF Synthesis for Examples from ISIC 2018 Dataset. We Find That, Even in a Multi-class Problem, Our Approach Is Able to Synthesize Concentrated Image Changes, Thus Enabling Us to Introspect Deep Models with Arbitrarily Complex Decision Boundaries. Moreover, Such Perturbations Are Consistent with the ABCD (Asymmetry, Border, Color and Diameter) Signatures Adopted by Clinicians for Diagnosing Lesions.

5.5.3 Choice of Strong Image Priors for Producing Discernible Changes in Counterfactuals

Although both DIP and INR are effective, a more rigorous comparison of those two image priors is required. For this purpose, we utilize two important evaluation metrics, namely: (i) ability to produce large discernible changes, measured using ℓ_2 error between $\mathbf{x}$ and $\bar{\mathbf{x}}$ in the pixel-space; and (ii) ability to produce *concentrated* image manipulations (C. Chang et al. 2018), measured by thresholding (< 0.05) the difference image $|\mathbf{x} - \bar{\mathbf{x}}|$ and determining the area of the largest bounding box that contains all the non-zero values (between 0 and 1). Ideally, explanations that have a larger MSE and consistently lower concentration are more likely to produce easily interpretable, localized changes. The results in Table 2 are obtained by randomly choosing 5000 images from the *non-smiling* class and synthesizing the corresponding counterfactuals for *smiling*. Similarly, for the ISIC2018 dataset, we used 800 randomly chosen images from the *MEL* class and generated CFs for *NEV*. The naïve *LSO* + *TV* + ℓ_2 baseline produces counterfactuals with a significantly large MSE as well as

Table 2. Evaluating the Quality of the Synthesized CFs on Celeba Faces and ISIC 2018 Skin Lesion Datasets. The MSE and Concentration Metrics for the Celeba Dataset Were Obtained Using CFs Synthesized for 5000 Images from the *Non-smiling* Class. On the Other Hand, for ISIC 2018, We Used 800 Images from the *MEL* Class and Generated CFs for Changing the Prediction to *NEV*.

Dataset	**Metric**	**Method**				
		LSO + TV + L2	**LSO + DIP**	**LSO + INR**	**LSO + DIP + DEP**	**LSO + INR + DEP**
CelebA	MSE	0.36 ± 0.18	**0.09 ± 0.05**	0.11 ±0.07	0.19 ±0.11	0.22 ±0.07
	Conc.	0.43 ±0.22	0.29 ±0.16	0.28 ± 0.19	0.26 ±0.15	**0.18 ± 0.11**
	CD	0.31 ± 0.05	0.23 ±0.03	0.26 ±0.06	0.15 ± 0.04	**0.08 ± 0.03**
ISIC	MSE	0.43 ± 0.17	**0.14 ± 0.08**	0.19 ±0.09	0.23 ±0.14	0.24 ±0.13
	Conc.	0.36 ±0.16	0.32 ±0.14	0.29 ± 0.13	0.25 ±0.09	**0.22 ± 0.07**
	CD	0.32 ±0.13	0.28 ±0.15	0.25 ±0.12	0.17 ± 0.05	**0.11 ± 0.01**

concentration score, indicating that the CFs are uninterpretable and contain irrelevant perturbations all over the image. As expected, incorporating a stronger prior improves the concentration significantly, while also being highly conservative in terms of MSE, *i.e.,* non-discernible changes. Finally, enforcing manifold consistency using DEP achieves an optimal trade-off between the two metrics and produces meaningful CFs (Figures 24 and 25). In particular, INR based generators produce highly concentrated image manipulations.

5.5.4 Components Required for Producing CFs with Low Classifier Discrepancy Scores

We now evaluate the quality of the counterfactuals using the *classifier discrepancy* (CD) score. For this purpose, we consider a random subset of 5000 images each from *non-smiling* (X_0) and *smiling* (X_1) classes respectively. Following the strategy described in Section 5.3.5, we train the classifiers $\mathcal{F}$ and $\mathcal{F}^c$, and measure the CD score as the difference in test accuracies, $Acc.(\mathcal{F}, X^{test}) - Acc.(\mathcal{F}^c, X^{test})$. Similarly,

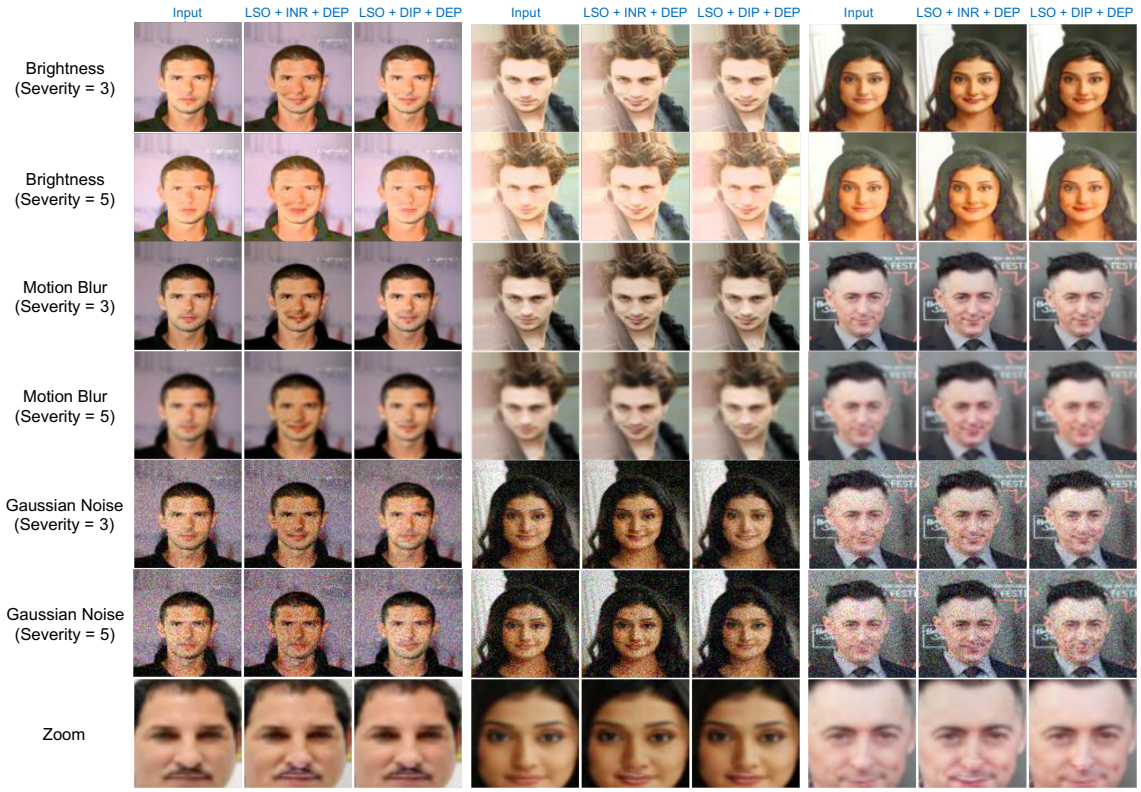


Figure 26. DISC Explanations Are Robust under Test-time Corruptions. We Find That Even under Unknown Test-time Corruptions, Our Approach Robustly Manipulates the Appropriate Regions in the Query Image (E.g., Mouth and Cheeks for Smiling). Such a Behaviour Can Be Attributed Both to the Ability of DEP to Reflect Challenging Distribution Shifts (JJ. Thiagarajan et al. 2021) and Our Progressive Optimization.

we repeat a simple evaluation for ISIC data by using 800 images from class *MEL* as X_0 and images from class *NEV* as X_1. From Table 2, we find that, using manifold consistency along with a strong image prior produces significantly lower CD scores (0.08 with *LSO + INR + DEP* on CelebA), when compared with LSO without manifold consistency (0.26 with *LSO + INR*). In particular, using INR + DEP fairs the best and consistently produces highly meaningful counterfactuals.

5.5.5 Robustness of Explanations Under Test-Time Distribution Shifts

In several practical applications, we often encounter shifts between the train and test distributions which makes model deployment challenging. Hence, we evaluate the behavior of our approach under unknown distribution shifts at test time. Note, we train the classifier on the clean CelebA faces dataset without introducing any corruptions. However, when we introduce corruptions at test-time, we find that (see Figure 26), our approach (*LSO + INR + DEP)* is able to robustly manipulate the appropriate regions in the query image. This can be attributed both to the ability of DEP to reflect challenging distribution shifts with higher error estimates (JJ. Thiagarajan et al. 2021) and our progressive optimization strategy to induce larger yet semantically meaningful changes (i.e., inherent noise clean up).

5.6 Summary

In this chapter, we developed DISC, a general approach to design counterfactual generators for any deep classifier, without requiring access to the training data or generative models. We drew connections to the problem of deep model inversion and extend it to support counterfactual generation. DISC can learn a CF generator on-the-fly by leveraging different image priors and manifold consistency constraints based on DEP, along with a progressive optimization strategy, to synthesize highly-plausible explanations. Future extensions to this work include exploring the use of multiple target attributes simultaneously in our optimization and applying this method to time-varying data.

Chapter 6

CALIBRATING CLASSIFIERS FOR IMPROVING ANOMALY DETECTION

In this chapter, we present the problem of anomaly detection under the multi-class classification setting. Although our approach is adaptable to any imaging modality, in this work we focus on medical anomaly detection. While there are many effective scoring functions for accurately identifying anomalies Liu et al. 2020, here, we highlight the importance of calibrating the detector using both inlier and outlier data for this application. We introduce the dual calibration objective, which allows the detector to maintain high accuracy for inliers while also rejecting examples from out-of-distribution regimes. We discuss the conventional methods for achieving this dual objective, including inlier specification using pixel space augmentations and outlier specification using a curated set of out-of-distribution data, along with their limitations. We show the performance of these existing methods on medical out-of-distribution detection to highlight the need for improved augmentation specifications. To that end, we find that inlier specification through augmentations in the latent space, along with exposure to diverse synthetic pixel space outliers derived from the training data, are essential for medical out-of-distribution detection. Through a rigorous empirical study on medical imaging benchmarks, we report significant performance gains (15% to 35% in AUROC) over existing approaches under both semantic and modality shifts.

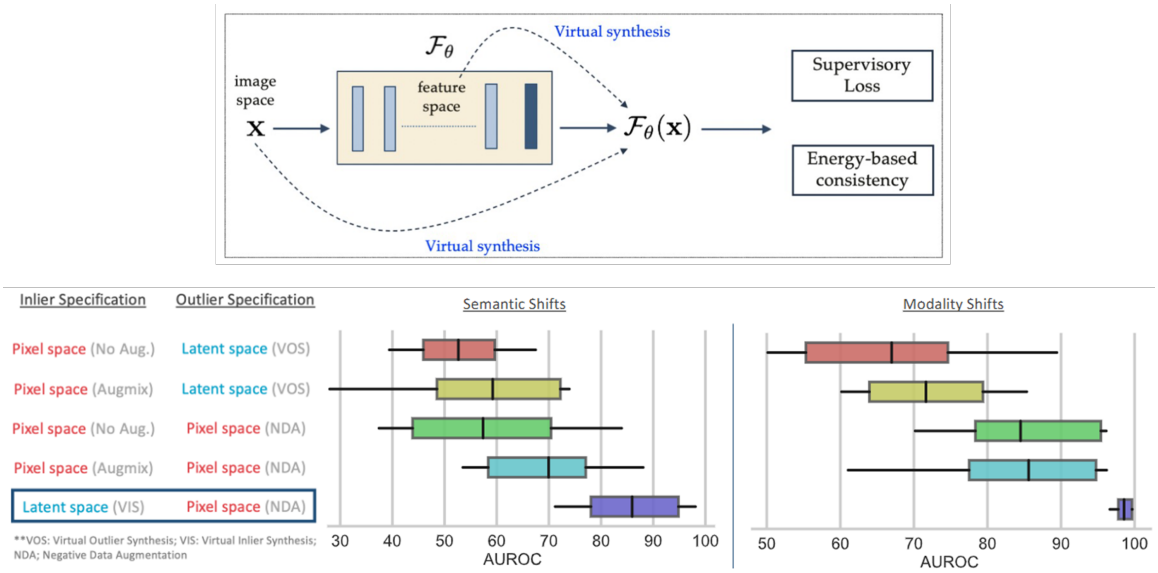


Figure 27. Specifying Synthetic Inliers/Outliers to Calibrate OOD Detectors. We Focus on Energy-based OOD Detectors for Deep Models and Explore the Design of Synthetic Augmentations. We Make a Striking Finding That the Space in Which the Different Augmentations Are Synthesized Plays a Critical Role on the Detection Performance. While State-of-the-art Approaches Such as VOS (Du et al. 2022) and NDA (Sinha et al. 2021) Can Fail Even in the Simpler Modality Change Detection (Far OOD), the Proposed Approach Consistently Leads to High-fidelity Detectors in Both Near and Far OOD Settings, Without Compromising the Test Accuracy.

6.1 Problem Setup

Detecting out-of-distribution (OOD) data characterized by a variety of semantic or covariate shifts with respect to the in-distribution (ID) data is vital for the safe adoption of AI tools in medical imaging (Hosny et al. 2018; Young et al. 2020). As a result, a broad class of inference-time, scoring functions (Hendrycks and Gimpel 2017; Liang, Li, and Srikant 2018; K. Lee et al. 2018; Liu et al. 2020) designed to distinguish between ID and OOD has emerged. However, in practice, one needs to calibrate those detectors, such that the dual objective of not compromising the test performance and reliably rejecting OOD data is effectively met.

Existing approaches for calibrating OOD detectors require users to specify regimes of inlier and outlier data. For example, a popular approach for specifying inliers is to leverage synthetic augmentations (Shorten and Khoshgoftaar 2019) that promote consistency and robustness under plausible ID data variations. Typical choices include geometric transforms such as rotation and translation (Wang, Perez, et al. 2017) or compositional strategies such as Augmix (Hendrycks et al. 2020), TrivialAug (Müller and Hutter 2021), Augmax (Wang et al. 2021), ALT (Gokhale et al. 2022) etc. On the other hand, Outlier Exposure (OE) (Hendrycks, Mazeika, and Dietterich 2018) that enforces the detector to not generalize to a carefully curated set of OOD data is the *modus operandi* for outlier specification. However, it is non-trivial to construct representative outlier set in practice. Hence, generating synthetic outliers in lieu of explicit curation is a viable alternative. For example, Du *et al.* (Du et al. 2022) synthesize outliers in the latent space of a classifier while Sinha *et al.* (Sinha et al. 2021) create image-space outliers as severely distorted, yet plausible, variations of the ID data with the aid of generative models. While they are well suited for natural image benchmarks, we make an interesting finding that OOD detectors constructed with existing choices for inlier/outlier specification are ineffective for both *Near OOD* and *Far OOD* settings (see Figure 27) in medical imaging.

In this paper, we posit that the space in which the inlier and outlier augmentations are specified plays a central role in improving the performance of medical OOD detectors. Using an energy based (Liu et al. 2020) training framework and a rigorous empirical study with benchmarks encompassing a wide-variety of modalities (skin lesions, histopathology slides, blood cells, tissues), dataset sizes, distribution shifts as well as architectural choices (WideResnet, ResNet-50 (He et al. 2016)), we show that inliers synthesized in the latent space coupled with diverse, image-space outliers

consistently produce high fidelity detectors. Note that, our approach is straightforward to be integrated with any prediction task or OOD scoring mechanism, and can produce significantly improved medical OOD detectors.

6.2 Related Work

6.2.1 Out-of-Distribution detection

Out-of-Distribution detection is the task of identifying whether a given sample is drawn from the in-distribution data manifold or not. Such a task requires an effective scoring metric that can well distinguish between ID and OOD data. In this context, much of recent research has focused on designing useful scoring functions to improve detection over different regimes of OOD data. For instance, Hendrycks *et al.* (Hendrycks and Gimpel 2017) proposed the Maximum Softmax Probability (MSP) score as a baseline method of OOD detection. Subsequently, Liang *et al.* (Liang, Li, and Srikant 2018) proposed ODIN which is a scoring function based on re-calibrating the softmax probabilities through temperature scaling and input pre-processing. On similar lines, Lee *et al.* (K. Lee et al. 2018) utilized Mahalanobis distances accumulated from the classifier latent spaces as a scoring metric. Ren *et al.* (Ren et al. 2021) proposed the relative mahalanobis distance as an effective score for fine grained OOD detection. Sastry *et al.* (Sastry and Oore 2020), proposed a latent space scoring metric for detecting outliers using Gram Matrices. More recently, Liu *et al.* (Liu et al. 2020) proposed using the energy metric as a scoring function for OOD detection. The metric is directly related to the underlying data likelihood and demonstrated to produce significant OOD detection improvements. Owing to the ease of adoption and success

of the energy metric in OOD detection, without loss of generality, we adopt energy as the scoring function in this paper.

6.2.2 OE-free OOD Detection

The objective defined in (6.2) requires the OOD detector to be calibrated with pre-specified curated outlier data. However, it is significantly challenging to construct such datasets in practice naturally motivating the design of 'OE-Free' methods. With the requirement of the ODIN detector to be fined tuned on pre-specified ID and OOD datasets, Hsu *et al.*(Hsu et al. 2020) proposed Generalized ODIN (G-ODIN) as an outlier data free method that adds on top of ODIN while significantly improving the detection performance. On the other hand, Du *et al.* (Du et al. 2022) synthesize virtual outliers by sampling hard negative examples (i.e, samples at the class decision boundaries) directly in the latent space of a classifier to calibrate the OOD detector in lieu of conventional pixel space methods without requiring specific outlier datasets. Our formulation broadly falls under this category as we synthesize outliers only using the training data without any external curation.

6.3 Preliminaries

6.3.1 Task Setup

We train a $K-$way classifier $\mathcal{F}_\theta$ using labeled data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$, where $\mathbf{x}_i$ is an image drawn from $P_{\text{ID}}(\text{x})$, and $y_i \in \mathcal{Y}_{\text{ID}} = \{1, 2, \cdots, K\}$ is its corresponding label. The goal of OOD detection is to flag samples $\bar{\mathbf{x}} \in P_{\text{OOD}}(\text{x})$ that may correspond to

covariate or semantic shifts respect to $P_{\text{ID}}(\text{x})$. We consider two prominent classes of outlier data: (a) *Near OOD* scenarios where $P_{\text{OOD}}(\text{x})$ and $P_{\text{ID}}(\text{x})$ share significant semantic concepts, but the set of target labels are different *i.e.*, $\mathcal{Y}_{\text{OOD}} \neq \mathcal{Y}_{\text{ID}}$. In medical imaging, this encompasses a variety of open-set recognition settings, namely novel disease states, healthy control groups and images of different organs acquired using the same modality (Yang, Zhou, and Liu 2022) (e.g., histopathology images of breast and colon cancer cells); (b) *Far OOD* scenarios where $P_{\text{OOD}}(\text{x})$ and $P_{\text{ID}}(\text{x})$ are disparate as well as $\mathcal{Y}_{\text{OOD}} \neq \mathcal{Y}_{\text{ID}}$. For instance, a chest X-ray image is an example of a far OOD sample for models trained on skin lesions.

6.3.2 Energy Based Framework for Medical OOD Detection

The free energy function for discriminative models (Liu et al. 2020) maps an input $\mathbf{x}$ to a deterministic scalar $E(\mathbf{x};\theta)$ that is linearly aligned with log-likelihood $\log(p_{\text{ID}}(\text{x}))$. Mathematically, $E(\mathbf{x};\theta) = -T\ \log\sum_{k=1}^{K}\exp\{\mathcal{F}_\theta^k(\mathbf{x})/T\}$, where $\mathcal{F}_\theta^k$ denotes the logit for class k and T is the temperature scaling parameter. We adopt the energy function to train an OOD detector G alongside the classifier, similar to (Liu et al. 2020), defined as follows:

$$G(\mathbf{x};\mathcal{F}_\theta,\tau) = \begin{cases} \text{outlier}, & \text{if } -E(\mathbf{x},\mathcal{F}_\theta) \leq \tau, \\ \text{inlier}, & \text{if } -E(\mathbf{x},\mathcal{F}_\theta) > \tau. \end{cases} \tag{6.1}$$

Here, τ is a user-defined threshold for detection. Since the training data is expected to be characterized by low energy in comparison to OOD, we use negative energy scores to align with the notion that ID samples should have higher scores over OOD samples.

In practice, it is important to calibrate G such that the dual objectives of not

compromising ID performance and reliably rejecting OOD are met. This can be formally stated as:

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{L}_{CE}(\mathcal{F}_\theta(\mathbf{x}), y) + \alpha \mathop{\mathbb{E}}_{\mathbf{x}\in\mathcal{D}_{\text{in}}} \mathcal{L}_{\text{ID}}(E(\bar{\mathbf{x}}); \theta) + \beta \mathop{\mathbb{E}}_{\bar{\mathbf{x}}\in\mathcal{D}_{\text{out}}} \mathcal{L}_{\text{OOD}}(E(\bar{\mathbf{x}}); \theta). \quad (6.2)$$

Here, $\mathcal{L}_{CE}(.)$ is the standard cross-entropy loss. The terms $\mathcal{L}_{\text{ID}}$ and $\mathcal{L}_{\text{OOD}}$ (implemented as margin losses) are used to calibrate the OOD detector to operate as expected in the regimes of the specified inliers ($\mathcal{D}_{\text{in}}$) and outliers ($\mathcal{D}_{\text{out}}$). The success of this optimization hinges on the appropriate specification of inliers and outliers, which is the focus of this work.

6.4 Approach

We study the implementation of (6.2) by exploring choices for inlier and outlier specification. In this context, we focus on the use of synthetic augmentations, without requiring additional data curation or explicit flagging of OOD data as in existing approaches.

6.4.1 Augmentations for Inlier Synthesis

A popular strategy for improving the generalization of classifier models is to leverage data augmentation strategies. While it is common to utilize pixel-space transformations, we propose to utilize feature space augmentations as an alternative choice for inlier synthesis.

6.4.1.1 Pixel-space Synthesis

These inliers are can be generated directly in the pixel-space by leveraging known statistical invariances of the image data. Following state-of-the-art, we consider the following strategies to perform inlier synthesis:- (i) conventional image manipulations such as random horizontal, vertical flips or color jitter and (ii) compositional strategies such as Augmix (Hendrycks et al. 2020) that synthesize inliers as a composition of multiple geometric and perceptual transformations.

Latent-space Synthesis While pixel-space augmentations are known to often aid the classifier performance, it has been shown that (Hendrycks et al. 2021) they can adversely impact model safety, e.g., outlier detection or calibration under real-world shifts due to over-generalization. In order to systematically calibrate OOD detectors, while also controlling the risk of over-generalization, we propose to synthesize inliers in the low-dimensional latent space of a classifier. Formally, we assume that the model $\mathcal{F}$ can be decomposed into feature extractor and classifier modules as $\mathcal{F} = h \circ c$ and we approximate data from class k in the feature space as $p(h(\mathbf{x})|y = k) \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}})$. Similar to (K. Lee et al. 2018), each class is modeled using a class-specific mean $\widehat{\boldsymbol{\mu}}_k \in \mathbb{R}^d$ and a shared covariance $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d \times d}$. Here, d denotes the latent feature dimension and the parameters are estimated using maximum likelihood estimation. In order to synthesize class-specific inliers, we sample each of the K gaussians from regions of low-likelihood corresponding to the tails as follows: $\mathcal{T} = \{\mathbf{t}_k | \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}) < \delta\}_{k=1}^{K}$. Here $\mathbf{t}_k$ denotes the inlier sampled from the k^{th} gaussian distribution. The modeling of class-specific gaussian distributions with a tied covariance allows the predictive model to be viewed under the lens of linear discriminant analysis (LDA) (K. Lee et al. 2018).

If $p(y|h(\mathbf{x}))$ denotes the inferred posterior label distribution, we have,

$$p(y=c|h(\mathbf{x})) = \frac{\exp\left(\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_c + \log\beta_c\right)}{\sum_{k=1}^{K}\exp\left(\widehat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k + \log\beta_k\right)}, \tag{6.3}$$

where β denotes the prior probabilities. On comparing (6.3) with the standard softmax based classifiers as well as with the definition of energy (Liu et al. 2020), we observe that,

$$E(\mathbf{x}, y=c) = -\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) + \frac{1}{2}\widehat{\boldsymbol{\mu}}_c^T \Sigma^{-1}\widehat{\boldsymbol{\mu}}_c - \log\beta_c. \tag{6.4}$$

Invoking the definition of the gaussian density function, and by expressing kernel parameters in terms of energy, we can relate the energy for the latent space mean $\widehat{\boldsymbol{\mu}}_k$ and tail $\mathbf{t}_k$ as

$$E\Big(h(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_k, y=k\Big) - E\Big(h(\mathbf{x}) = \mathbf{t}_k, y=k\Big) < \frac{1}{2}(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}_k + \widehat{\boldsymbol{\mu}}_k). \tag{6.5}$$

For simplicity, we reuse the same notation E to define the energy for $\mathbf{x} \in \mathcal{D}$ or equivalently $h(\mathbf{x})$ in the latent space. We find that the free energy $E(h(\mathbf{x}) = \mathbf{t}_k)$ can be bounded as:

$$E(h(\mathbf{x}) = \mathbf{t}_k) > -\log\sum_{k=1}^{K} \exp\Big(- E(h(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_k, k) + \frac{1}{2}(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}_k + \widehat{\boldsymbol{\mu}}_k)\Big) \tag{6.6}$$

Our optimization in (6.2) attempts to minimize the free energy for the inlier samples $\mathbf{t}_k$. From the expression (6.6), it becomes apparent that the model is encouraged to minimize the term $(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)$, *i.e.*, push the tail samples closer to the class-specific means and thereby improve generalization beyond the prototypical samples. When compared to pixel-space inliers, our approach includes more challenging examples,

albeit with reduced diversity. From our empirical study, we find that, when combined with an appropriate outlier specification, this leads to significant improvements in the challenging Near OOD detection.

6.4.2 Augmentations for Outlier Synthesis

In addition to inlier specification, exposure to representative outliers (Hendrycks, Mazeika, and Dietterich 2018; Roy et al. 2022; Thulasidasan et al. 2021; Sinha et al. 2021; J. Zhang et al. 2021; Chen et al. 2021) is critical to calibrate OOD detectors. Since carefully curated, diverse outlier datasets are not always available, we resort to generating synthetic outliers using the training data either in the latent or pixel-space.

6.4.2.1 Latent-space Synthesis

Following (Du et al. 2022), we synthesize latent-space outliers as tail samples from class-specific gaussians in the penultimate layer of a classifier. Using (6.2), we enforce such samples to be characterized with maximum free energy.

6.4.2.2 Pixel-space Synthesis

In our approach, we propose to utilize pixel-space outliers as a set of severely corrupted versions of training samples. This is motivated by the need for exposing models to rich outlier data, so that the OOD detector can be calibrated to handle a variety of OOD scenarios. In contrast to latent-space outliers, pixel-space outliers distort the global features of the training distribution and produces statistically

disparate examples. In our implementation, we consider two augmentation strategies, where one of them is randomly chosen in every iteration: (i) `Augmix o Jigsaw`: We first transform an image using Augmix (Hendrycks et al. 2020) with high severity (set to 11), and subsequently distort using the Jigsaw corruption (divide an image into 16 patches and perform patch permutation); (ii) `RandConv` (Xu et al. 2021): We used random convolutions with very large kernel sizes (chosen from $9 - 19$) to produce severely corrupted versions of the training images. We find that the inherent diversity of this outlier construction consistently leads to large performance gains, in particular for Far OOD, in comparison to latent-space outliers which offer limited diversity.

6.4.3 Training

To implement out approach, we define the loss functions in (6.2) as follows:

$$\mathcal{L}_{\text{ID}} = \left[\max\left(0, E(h(\mathbf{x}) = \mathbf{t}_k) - m_{\text{ID}}\right)\right]^2; \mathcal{L}_{\text{OOD}} = \left[\max\left(0, m_{\text{OOD}} - E(\mathbf{x} = \bar{\mathbf{x}})\right)\right]^2.$$

Here, $\mathcal{L}_{\text{ID}}$ is a margin based loss with margin parameter m_{ID} (set to -20) for minimizing the energy $E(.)$ of the synthesized inliers. Similarly, for the outlier data, we define $\mathcal{L}_{\text{OOD}}$ with margin parameter m_{OOD} (set to -7), so that the energy for those samples is maximized.

6.5 Experiment Setup

6.5.1 ID Datasets.

In this paper, we use a large suite of medical imaging benchmarks of varying dataset sizes and image resolutions to evaluate our proposed approach.

6.5.1.1 MedMNIST Benchmark

(i) BloodMNIST consists of $17,092$ human blood cell images collected from healthy individuals corresponding to 8 different classes; (ii) PathMNIST is a histology image dataset of colorectal cancer with $107,180$ samples of non-overlapping, hematoxylin and eosin stained image patches from 9 different classes; (iii) DermaMNIST is a skin lesion dataset curated from the HAM1000 (Tschandl *et al.*) database. It contains a total of $10,015$ images across 7 cancer types; (iv) OctMNIST contains $109,309$ optical coherence tomography (OCT) retinal images corresponding to 4 diseases; (v) TissueMNIST is a kidney cortex image dataset curated from the Broad Bioimage Benchmark Collection with $236,386$ images from 8 classes; (vi) (vii) (viii) Organ(A,C,S)MNIST are images of abdominal CT collected from the Axial, Coronal and Sagittal planes of 3D CT images from the Liver-tumor segmentation benchmark. The datasets contain $58,850$, $23,660$ and $25,221$ images across 11 classes respectively.

6.5.1.2 ISIC2019 Skin Lesion Dataset

(Tschandl, Rosendahl, and Kittler 2018; Codella et al. 2018; Combalia et al. 2019) is a skin lesion classification dataset containing a total of $25,331$ images belonging to 8 disease states namely Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC) and Squamous cell carcinoma (SCC).

6.5.1.3 NCT (Colorectal Cancer)

(Kather, Halama, and Marx 2018) contains $100,000$ examples of 224×224 histopathology images of colorectal cancer and normal tissues from 9 possible categories namely, Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

6.5.2 Out-of-Distribution Datasets

In case of the MedMNIST benchmark, for every dataset i, we consider the validation splits of each of the remaining datasets ($j \neq i$) for modality shift detection. On the other hand, we only consider the classes unseen during training to evaluate semantic shift detection.

For the high resolution image benchmarks (ISIC2019 and NCT (Colorectal)), the following datasets are used to evaluate OOD detection under modality (M: ID) and semantic shifts (S: ID):- (i) Camelyon-17 (WILDS)(M: ISIC,S: NCT) is a histopathology dataset of tumor and non-tumor breast cells with approximately 450K images curated from five different medical centers. We randomly sample 3000 examples from the dataset for OOD detection; (ii) Knee (M: ISIC,M: NCT) Osteoarthritis severity grading dataset contains X-ray images of knee joints with examples corresponding to arthritis progression. We use 825 examples chosen randomly from the dataset for evaluation; (iii) CXR (M: ISIC,M: NCT)[1] is a chest X-ray dataset curated from the MIMIC-CXR database containing $1,083$ samples corresponding to disease states

[1]https://github.com/cxr-eye-gaze/eye-gaze-dataset

namely normal, pneumonia and congestive heart failure and (iv) Retina (M: ISIC,M: NCT) is a subset of 1500 randomly chosen retinal images with different disease progressions from the Diabetic Retinopathy detection benchmark from Kaggle[2]; (v) Clin Skin (S: ISIC)(Pacheco *et al.*)contains 723 images of healthy skin; (vi) Derm-Skin (S: ISIC)(Pacheco *et al.*) consists of 1565 dermoscopy skin images obtained by randomly cropping patches in the ISIC2019 database; (vii) NCT 7K (S: NCT)(Pacheco *et al.*) contains 1350 histopathology images of colorectal adenocarcinoma with no overlap with NCT. In addition, we use 2000 randomly chosen examples from ISIC as a source of modality shift for the detector trained on NCT and vice-versa. Moreover, novel classes unseen while training are also used to evaluate detection under semantic shifts.

6.5.3 Evaluation Metrics

(i) Area Under the Receiver Operator Characteristic curve (**AUROC**), a threshold independent metric, reflects the probability that an in-distribution image is assigned a higher confidence over the OOD samples; (ii) Area under the Precision-Recall curve (**AUPRC**) where the ID and OOD samples are considered as positives and negatives respectively.

[2]https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data

6.6 Experiment Details

6.6.1 Dataset Preprocessing

We first split each of the datasets into two categories namely (i) data from classes known during training (*known classes*) and (ii) data from classes unknown while training (*novel classes*) where the latter constitutes OOD data with semantic shifts. The dataset from the former category is split in the ratio of $90 : 10$ for training and evaluating the predictive models. Table 3 provides the list of known and novel classes for the MedMNIST benchmark. In case of ISIC2019, we choose BKL, VASC and SCC as novel classes while MUC, BACK and NORM are chosen as novel classes for the NCT(Colorectal) benchmark. In both cases, we utilize the remaining classes for training and evaluating the respective detectors.

6.6.2 Choice of OOD Detector Architecture

For all experiments with the MedMNIST benchmark, we resize the images to 32×32 and utilize the $40 - 2$ WideResNet architecture. On the other hand, for experiments on ISIC2019 and NCT, we resize the images to 224×224 and employ the ResNet-50 model pre-trained on imagenet.

6.6.3 Training Details

6.6.3.1 Estimating Class-specific Means and Joint Covariance

We estimate the means and joint covariance via maximum likelihood estimation during training, similar to Du *et al.* We employ K queues each of size 1000 where each queue is filled during every iteration until their pre-specified capacities with the class specific latent embeddings (extracted from the penultimate layer) of the training data. We then adopt an online strategy to update the queues such that they contain much higher quality embeddings of the data as the training progresses. In particular, we enqueue one class-specifc latent embedding to the respective queues while dequeuing one embedding from the same class.

6.6.3.2 Sampling the Latent Space

In practice, we select samples close to the class specific boundaries based on the n^{th} smallest likelihood ($n = 64$) among N examples ($N = 10,000$) synthesized from the respective Gaussian distributions.

6.6.3.3 General Hyper-parameters

We train the $40-2$ WideResNet and ResNet-50 architectures for 100 and 50 epochs with learning rates of $1e{-}3$ and $1e{-}4$ respectively. We reduce the learning rate by a factor of 0.5 every 10 epochs using the Adam optimizer with a momentum of 0.9 and a weight decay of $5e^{-4}$. We choose a batch size of 128 for datasets from MedMNIST

Table 3. Known and Novel Classes Selected From the MedMNIST Benchmark

Datasets	**Blood**	**Path**	**Derma**	**OCT**	**Tissue**	**OrganA,C,S**
Known Classes	$1-5,7$	$1-5,7$	$1,3-6$	$1,2,4$	$1-2,4-5,7-8$	$1,5-11$
Novel Classes	$6,8$	$6,8,9$	$2,7$	3	$3,6$	$2,3,4$

Table 4. Modality Shift Detection on the MedMNIST Benchmark. We Report Detection Accuracies Obtained Using Different Approaches with a $40-2$ WideResnet Backbone. Note, for Each ID Dataset, We Show the Mean and Standard Deviation of AUROC Scores from Multiple OOD Datasets. In Each Case, the First and Second Best Performing Methods Are Marked in Green and Orange Respectively.

In Dist.	**Methods**					
	G-ODIN	**VOS**	**Aug. + VOS**	**NDA**	**Aug. + NDA**	**Ours**
Blood	88.7 ± 10.5	89.4 ± 12.9	84.2 ± 11.6	96.2 ± 9.1	95.8 ± 5.5	99.7 ± 0.5
Path	84.4 ± 9.2	77.5 ± 10.7	71.0 ± 14.8	96.1 ± 4.2	61.1 ± 12.1	98.9 ± 1.7
Derma	85.3 ± 8.2	64.1 ± 16.2	85.3 ± 2.8	95.2 ± 5.4	80.0 ± 11.8	96.6 ± 4.2
OCT	49.0 ± 28.0	50.1 ± 5.8	68.0 ± 4.0	92.8 ± 16.8	94.4 ± 3.8	99.6 ± 0.9
Tissue	82.7 ± 28.3	72.9 ± 12.8	60.2 ± 16.6	81.1 ± 30.6	70.2 ± 32.1	96.6 ± 8.4
OrganA	95.8 ± 3.0	73.7 ± 13.1	77.8 ± 11.9	70.2 ± 21.9	96.2 ± 5.0	99.7 ± 0.4
OrganS	80.3 ± 25.3	51.5 ± 17.1	62.1 ± 11.4	94.0 ± 4.2	92.9 ± 11.6	98.2 ± 3.1
OrganC	85.7 ± 6.9	56.6 ± 3.6	64.6 ± 14.4	93.2 ± 4.4	94.2 ± 8.5	99.1 ± 0.6

and 64 for the full-sized images. For all experiments including the baselines (except G-ODIN), we use a margin $m_{\text{ID}} = -20$ and $m_{\text{OOD}} = -7$ with $\alpha = \beta = 0.1$. We introduce NDA during the beginning of training for our approach and baselines except for the VOS variants where we introduce the outliers at epoch 40 following standard practice.

Table 5. Semantic Shift Detection on the MedMNIST Benchmark. We Report AUROC Scores for Detecting Novel Classes Using Different Approaches with a $40-2$ WideResnet Backbone.

In Dist.	Methods					
	G-ODIN	**VOS**	**Aug. + VOS**	**NDA**	**Aug. + NDA**	**Ours**
Blood	53.91	44.66	38.16	65.96	53.5	89.15
Path	51.75	39.41	71.74	37.43	56.99	71.2
Derma	69.34	67.46	72.86	51.23	69.94	75.55
OCT	47.19	55.37	52.0	50.98	75.2	78.9
Tissue	55.17	46.37	27.98	42.29	58.83	83.37
OrganA	89.86	62.19	73.92	44.41	75.59	98.1
OrganS	81.96	46.98	71.99	83.95	88.08	93.95
OrganC	79.32	58.77	65.17	83.76	81.49	97.46

Table 6. Evaluation on the ISIC 2019 Benchmark. We Report AUROC Scores Obtained with a Resnet-50 Model Trained on the ISIC 2019 Dataset. Note, We Show Results for Both Semantic Shifts (Blue) and Modality Shifts (Red).

OOD Data	Methods					
	G-ODIN	**VOS**	**Aug. + VOS**	**NDA**	**Aug. + NDA**	**Ours**
Novel Classes	62.2	75.04	68.69	65.41	68.38	74.0
Clin Skin	62.93	61.33	78.8	67.06	72.01	81.55
Derm Skin	71.93	80.53	79.27	82.73	85.5	93.9
Wilds	65.78	66.71	57.15	83.69	85.29	99.77
Colorectal	77.08	32.02	81.27	71.33	78.84	98.58
Knee	66.5	23.02	83.47	89.25	94.73	94.08
CXR	74.32	76.8	80.93	83.18	62.08	96.94
Retina	71.1	76.33	87.39	76.04	76.65	95.86
Avg.	68.98	61.47	77.12	77.34	77.94	**91.84**

Table 7. Evaluation on the Colorectal Cancer Benchmark. We Report AUROC Scores Obtained with a Resnet-50 Model Trained on the the Colorectal Cancer Dataset (Kather, Halama, and Marx 2018). Note, We Show Results for Both Semantic Shifts (Blue) and Modality Shifts (Red).

OOD Data	**Methods**					
	G-ODIN	**VOS**	**Aug. + VOS**	**NDA**	**Aug. + NDA**	**Ours**
Novel Classes	41.59	84.24	63.13	79.38	74.34	94.06
NCT 7K	76.02	78.92	62.04	80.46	63.25	96.11
WILDS	43.82	95.97	87.31	42.73	79.4	92.47
ISIC2019	79.03	65.6	85.46	98.71	65.17	99.86
Knee	95.55	95.26	58.87	96.63	44.67	99.98
CXR	95.99	99.19	67.18	99.79	71.65	99.91
Retina	96.67	81.06	95.62	99.68	54.66	100.0
Avg.	75.52	85.75	74.23	85.34	64.73	**97.48**

6.7 Findings

6.7.1 Modality Shift Detection on MedMNIST

In this study, we compare the OOD detection performance of the proposed approach against the baselines across the 8 benchmarks from MedMNIST. All OOD detection methods for this benchmark were designed based on a $40-2$ WideResNet feature extractor backbone (Zagoruyko and Komodakis 2016). In each experiment, one of the 8 datasets was considered as ID and the modality shift detection performance was evaluated using the remaining datasets. In Table 4, we report the mean AUROC scores and standard deviations for all detection approaches and benchmarks (fine-grained results and additional metrics are provided in the appendix). It can be observed

that the proposed approach consistently outperforms the baselines by significant margins while exhibiting low variance in detection performance across benchmarks. Interestingly, state-of-the-art baselines such as G-ODIN and VOS under-perform on these real-world benchmarks, thus emphasizing the importance of exposing the detector to diverse negative examples. This observation is further validated by the effectiveness of the NDA baseline over VOS, which predominantly generates hard negatives.

6.7.2 Semantic Shift Detection on MedMNIST

In this experiment, we evaluate the ability of our approach in recognizing semantic shifts with respect to the underlying training distribution. For each of the MedMNIST benchmarks, we held-out a subset of classes during training, which are considered as semantic shifts. The inherently homogeneous nature of medical images and subtle variations in the image statistics across different classes (Cao, Hui, et al. 2020) makes this very challenging. Furthermore, improving the sensitivity of OOD detectors in this setting will enable models defer from making an incorrect diagnosis. In Table 5, we report the AUROC scores for the semantic shift detection task with the same $40 - 2$ WideResNet backbone. Energy-based detectors designed with our approach produce the best AUROC scores across all datasets (except for the case of PathMNIST). While the G-ODIN detector (with learned additive noise magnitude), performs competitively in some cases, albeit demonstrating a large variance across benchmarks.

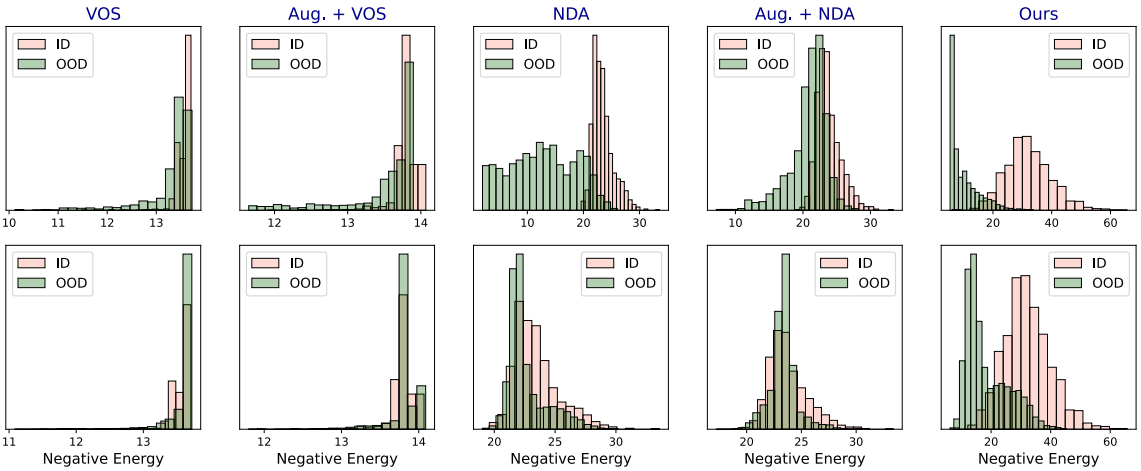


Figure 28. Histograms of Negative Energy Scores. We Plot the Scores Obtained Using Different Inlier and Outlier Specifications. With Blood MNIST as ID, the Top Row Corresponds to Modality Shift (OOD: Derma MNIST) and the Bottom Row Shows Semantic Shift (OOD: Novel Classes).

6.7.3 Choice of Detector Architecture and Image Resolution

Next, we perform rigorous evaluations on ISIC2019 Skin Lesion and colorectal cancer histopathology benchmarks, which contain higher resolution images (224×224) compared to MedMNIST. Further, we also vary the architecture of the backbone (Resnet-50 (He et al. 2016)) to study the generality of our method. Similar to the previous study, we consider a large suite of semantic and modality shifts, and evaluate the performance using the AUROC metric (additional evaluations are included in the supplement). Tables 6 and 7 show the results for ISIC2019 and colorectal cancer benchmarks across different OOD settings. In each case, the OOD scenarios are appropriately categorized into semantic (blue) and modality (red) shifts respectively. It can be observed that, in the case of ISIC2019, our approach improves upon state-of-the-art methods, namely G-ODIN and VOS, by margins of 22% and 13% respectively.

Interestingly, on the colorectal cancer benchmark, NDA achieves detection accuracies that are comparable to our approach, particular in the case of modality shifts.

6.7.4 Discussion

From the empirical results in this study, we observe that introducing pixel space diversity via negative data augmentation is critical for accurately detecting modality shifts. For instance, NDA & Aug. + NDA produce improved detection scores over VOS and Aug. + VOS baselines, which synthesize latent space outliers with limited diversity. On the other hand, when detecting semantic shifts, we find that VOS and Aug.+VOS typically provide significant boosts over NDA and G-ODIN. We find that such a strategy which samples hard outliers in the latent space can be useful for improving the detector's sensitivity to near-OOD samples. Overall, by explicitly controlling ID generalization using latent inlier synthesis and exposure to diverse synthetic outliers via NDA, our approach produces much higher quality OOD detectors. Chapter 28 depicts the histograms of the negative energy scores for the case of BloodMNIST (ID), wherein the modality shift results were obtained using DermaMNIST and the semantic shift corresponds to novel classes. We observe that our approach effectively distinguishes between ID and OOD distributions (much higher scores for ID data) in both cases, as illustrated by well-separated distributions, while the other approaches contain high overlap in their scores.

6.8 Summary

In this chapter, we presented the problem of anomaly detection with a focus on medical imaging. We emphasized the importance of calibrating the detector using both inlier and outlier data, and introduced a dual calibration objective which allows the detector to maintain high accuracy for inliers while also rejecting examples from out-of-distribution regimes. We showed the performance of different existing methods on medical out-of-distribution detection and identified that the space in which the inliers and outliers are specified is critical to improve OOD detection. We found that inlier specification through augmentations in the latent space, along with exposure to diverse synthetic pixel space outliers derived from the training data, are essential for medical out-of-distribution detection. Such a training strategy can help avoid models from being blindsided when deployed in the open-world and guide practitioners in decision making.

Chapter 7

Δ-UQ - DESIGNING SINGLE MODEL UNCERTAINTY ESTIMATORS VIA STOCHASTIC DATA CENTERING

In this chapter, we take a step forward from deterministic models in an effort to rigorously characterize model confidences. Identifying the potential sources of error that can arise during the training pipeline can provide useful signals of model failure guiding practitioners and researchers. Broadly referred to as Uncertainty Quantification (UQ), incorporating the same during model training allows the DNN prediction to be a distribution in lieu of conventional point estimates. While uncertainties can be broadly classified as *aleatoric* (irreducible noise in the data) and *epistemic* (model errors), there have been recent efforts for e.g., Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) towards reliably estimating such uncertainties. However, existing methods offer significant challenges such as poor approximations of predictive posteriors and computational complexity. We present a principled finding that an ensemble of neural networks with the same weight initialization, trained on datasets that are shifted by a constant bias gives rise to slightly inconsistent trained models, where the differences in predictions are a strong indicator of epistemic uncertainties. Since this is achieved via a trivial input transformation, we further show that it can therefore be approximated using just a single neural network. We systematically show that our uncertainty estimates are superior in a wide variety applications and benchmarks and provide key observations.

7.1 Problem Setup

Accurately estimating uncertainties in DNNs is an active area of research due to its implications in a wide range of scientific and engineering problems. Broadly, there are two kinds of uncertainties that are often considered – (a) *aleatoric*: uncertainties in the data generating process that are typically irreducible, and (b) *epistemic*: uncertainties of the model that can be reduced by observing more data, which is the focus of our work. Some of the popular techniques for the latter include Bayesian methods (Wilson and Izmailov 2020; He, Lakshminarayanan, and Teh 2020; Neal 2012; Blundell et al. 2015) that use a prior on the network weights, Monte Carlo approximations such as MC Dropout (Gal and Ghahramani 2016) that approximate sampling from the posterior, and empirical methods such as Deep Ensembles (DEns) (Lakshminarayanan, Pritzel, and Blundell 2017). In particular, DEns trains an ensemble of neural networks with different initializations, such that the uncertainty estimate on a test sample is given by the inconsistency between predictions from the member models. In practice, DEns has been found to often outperform other related methods (Ovadia et al. 2019; Lakshminarayanan, Pritzel, and Blundell 2017; Van Amersfoort et al. 2020), but it comes at the cost of training several DNNs to obtain reliable uncertainties (typically $10 - 20$), which is a severe computational bottleneck when it comes to modern DNNs especially on large scale datasets. In light of this, there is increased interest in developing single model estimators that can still produce high quality uncertainties. There has been some promising work in this direction in the recent past, specific to deep classifiers (Van Amersfoort et al. 2020) or regression models (Jain et al. 2021), and has been shown to perform comparably to DEns in some use-cases.

In this work, we first begin by exploring an alternate approach for constructing deep

ensembles – instead of using multiple randomized weight initializations, we propose to shift the training domain using random biases, such that every model in the ensemble is trained with data that has been shifted by a different *constant* bias c. Formally, for a labeled dataset $\{(\mathrm{x}, y)\}$, the k^{th} model is trained to fit $\{(\mathrm{x} - \mathrm{c}_k, y)\}$, where c_k (referred as an anchor) is of the same size as x. Though this manipulation appears trivial, the kernel induced by deep networks is not inherently shift invariant(Jacot, Gabriel, and Hongler 2018; Tancik et al. 2020), thus implying each DNN learns a slightly different model due to the bias. This leads to one of our key observations:

> **Anchor Ensemble:** *When an ensemble of DNNs, with the same fixed initialization, are trained on a dataset shifted by random constant biases, the variation across the ensemble's predictions is a strong indicator of model uncertainty.* [1]

Based on analysis with the neural tangent kernel (NTK) (Jacot, Gabriel, and Hongler 2018), we show that when the anchor c is made a random variable, the effective kernel is stochastic such that for each c, the model converges to a slightly different kernel. Consequently, this anchoring-based ensembling provides a different approach to DEns for sampling solutions from the hypothesis space. However, interestingly, our ensembling lends itself easily to be approximated using a single DNN:

> $\Delta-$**UQ** : *For a random anchor* c, [1] *can be approximated using a single DNN trained on the dataset transformed as* $\{\mathrm{x}, y\} \rightarrow \{[\mathrm{c}, \mathrm{x} - \mathrm{c}], y\}$. [2]

During inference, we obtain multiple predictions for a given sample by varying the choice of the anchor, such that the standard deviation of the predictions is our estimate for uncertainty. Without affecting the performance of the model, we find that $\Delta-$UQ produces meaningful epistemic uncertainties that we validate in a variety

of applications: outlier rejection, calibration under distribution shift on ImageNet, and sequential optimization of a large suite of black-box functions. We observe that $\Delta-$UQ consistently outperforms existing uncertainty estimates, while also being efficient to train as a single model estimator. With just a few simple changes, one can easily modify the training of any existing DNN model to support $\Delta-$UQ estimation. We include a Torch implementation in the supplement, and will make our code public for easy benchmarking.

7.2 Notations

Denote training data as $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^n$, where $\mathrm{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, to train a neural network $f(\boldsymbol{\theta}) \in \mathcal{H}$ with randomly initialized weights $\boldsymbol{\theta}_0$, such that a loss function $\mathcal{L}$ is minimized, *i.e.*, $\arg\min_{\boldsymbol{\theta}} \mathcal{L}(f(\mathrm{x}; \boldsymbol{\theta}), y)$. Here, $\mathcal{H}$ denotes the hypothesis space of potential solutions for fitting the observed data. Given a prior distribution on the weights $p(\boldsymbol{\theta})$, we can define the posterior over $\boldsymbol{\theta}$ as $p(\boldsymbol{\theta}|\mathcal{D})$ and subsequently, infer the posterior predictive distribution for a test sample (x_t, y_t). This can be used to quantify the uncertainty around the prediction as $p(y_t|\mathrm{x}_t, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y_t|\mathrm{x}_t, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$.

7.3 Related Work

A challenge for neural networks, however, is that the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ is computationally intractable. This has motivated the use of *Bayesian Neural Networks* (BNNs) (Neal 2012) using different approximations to the posterior including Monte-Carlo Dropout (Gal and Ghahramani 2016), variational inference (Graves 2011; Blundell et al. 2015), and sampling methods such as Markov chain Monte Carlo (Neal

2012; Welling and Teh 2011). In parallel, it has been empirically shown that Deep Ensembles (DEns) (Lakshminarayanan, Pritzel, and Blundell 2017) often tend to outperform Bayesian methods in terms of model calibration performance, even under challenging distribution shifts (Ovadia et al. 2019).

The success of DEns has been attributed to its ability to sample different functional modes (Fort, Hu, and Lakshminarayanan 2019) from the hypothesis space, and thus approximate the posterior predictive distribution (Wilson and Izmailov 2020). However, a critical limitation of DEns is the need to train multiple models (typically $10 - 20$) in order to obtain well calibrated uncertainties, which can be impractical for complex model architectures that have become commonplace today.

Characterizing the behavior of deep uncertainty estimators has mostly been done using empirical evaluation based on model calibration or out-of-distribution detection, but the recent advances in the neural tangent kernel (NTK) theory (Jacot, Gabriel, and Hongler 2018; Arora et al. 2019; Bietti and Mairal 2019; J. Lee et al. 2019) provide a convenient framework for more rigorous analysis. The basic idea of NTK is that, when the width of a neural network tends to infinity and the learning rate of SGD tends to zero, the function $f(\mathrm{x}; \boldsymbol{\theta})$ converges to a solution obtained by kernel regression using the NTK defined as $\mathbf{K}_{\mathrm{x}_i\mathrm{x}_j} = \mathbb{E}_{\boldsymbol{\theta}} \left\langle \frac{\partial f(\mathrm{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\mathrm{x}_j, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle$.

When the samples $\mathrm{x}_i, \mathrm{x}_j \in \mathcal{S}^{d-1}$, i.e., points on the hypersphere and have unit norm, the NTK for a simple 2 layer ReLU MLP can be simplified as a dot product kernel (Arora et al. 2019; Bietti and Mairal 2019; J. Lee et al. 2019):

$$\mathbf{K}_{\mathrm{x}_i\mathrm{x}_j} = h_{\mathrm{NTK}}(\mathrm{x}_i^\top \mathrm{x}_j) = \frac{1}{2\pi} \mathrm{x}_i^\top \mathrm{x}_j (\pi - \cos^{-1}(\mathrm{x}_i^\top \mathrm{x}_j)) \tag{7.1}$$

The NTK framework and its extensions that enable a posterior interpretation in the infinite limit have been used to study deep ensembles (He, Lakshminarayanan, and Teh 2020). Previous work (J. Lee et al. 2018; G. Matthews et al. 2018; Novak

et al. 2019) has also shown the existence of a distinct yet related kernel, referred to as the neural network Gaussian process (NNGP) kernel, where the initialization tends to a GP in infinite width limit.

7.4 Uncertainty Estimation with $\Delta-$UQ

As discussed above, most existing frameworks sample the hypothesis space either through different random initializations of ($\boldsymbol{\theta_0}$), perturbing the weight space after training, or using Bayesian methods by placing a prior on $p(\boldsymbol{\theta})$. Here, we propose to use a new strategy that involves injecting multiple trivial biases into the training data and analyze the resulting models using the NTK framework.

7.4.1 Anchor Ensembles: Ensembling by Injecting Trivial Biases

Let us examine the scenario where we shift an entire dataset (both train and validation) using a constant bias, c, to obtain a new dataset $\mathcal{D}_{\mathrm{c}}$ using which we train the model f_{c}. Since we always choose c from the training distribution at random, this has the effect of zero-centering the dataset around different training points. Finally, we are interested in the relationship between the models $\{f_{\mathrm{c}_1}, f_{\mathrm{c}_2}, \ldots, f_{\mathrm{c}_k}\}$. If the NTK induced by f is shift-invariant (for e.g., when Fourier features (Tancik et al. 2020) are used), the shifts make no difference, resulting in identical models $f_{\mathrm{c}_1} = \cdots = f_{\mathrm{c}_k}$. However, since NTKs for models like MLPs and CNNs are not inherently shift-invariant (J. Lee et al. 2019), we find that the models lead to an effective deep ensemble, wherein the variation across the predictions is a strong indicator of epistemic uncertainties.

(a) Effect of shifted training on NTK: We are interested in understanding how (7.1)

changes when the *entire training domain* is shifted by c – i.e., $h_{\text{NTK}}((\mathrm{x}_i - \mathrm{c})^\top(\mathrm{x}_j - \mathrm{c}))$. Without loss of generality, for the sake of notational convenience, we assume $\mathrm{x}_i - \mathrm{c}$ and $\mathrm{x}_j - \mathrm{c}$ are also made unit norm. To simplify the expansion, we use a Taylor series expansion for the $\cos^{-1}$ function: $\cos^{-1}(\mathrm{u} - \mathrm{c}) \approx \cos^{-1}(\mathrm{u}) + \frac{\mathrm{c}}{\sqrt{1-(\mathrm{u}-\mathrm{c})^2}}$.

Expanding $(\mathrm{x}_i - \mathrm{c})^\top(\mathrm{x}_j - \mathrm{c})$ as $\mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top(\mathrm{x}_i + \mathrm{x}_j - \mathrm{c})$ and letting $\mathrm{v} = (\mathrm{x}_i + \mathrm{x}_j - \mathrm{c})$, we obtain the expression for h_{NTK} under a shifted domain as follows:

$$\begin{aligned}
&\mathbf{K}_{(\mathrm{x}_i-\mathrm{c})(\mathrm{x}_j-\mathrm{c})} = \frac{1}{2\pi}(\mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top \mathrm{v})(\pi - \cos^{-1}(\mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top \mathrm{v})) \\
&\approx \frac{1}{2\pi}\mathrm{x}_i^\top \mathrm{x}_j(\pi - \cos^{-1}(\mathrm{x}_i^\top \mathrm{x}_j)) - \frac{1}{2\pi}\mathrm{c}^\top \mathrm{v}(\pi - \cos^{-1}(\mathrm{x}_i^\top \mathrm{x}_j)) - \frac{\mathrm{c}(\mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top \mathrm{v})}{2\pi\sqrt{1 - (\mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top \mathrm{v})^2}} \\
&= \mathbf{K}_{\mathrm{x}_i\mathrm{x}_j} - \Gamma_{\mathrm{x}_i,\mathrm{x}_j,\mathrm{c}},
\end{aligned} \tag{7.2}$$

where we combine all terms dependent on c into $\Gamma_{\mathrm{x}_i,\mathrm{x}_j,\mathrm{c}}$, which also behaves as a dot product kernel. From (7.2), we note that a trivial shift in the domain results in a non-trivial shift in the NTK function itself. In other words, (7.2) outlines the *effective* NTK as a function of c. We also note that $\mathbf{\Gamma}$ does not affect the spectral properties of the original NTK, as we observe in Figure 29.

Let us now consider the prediction on a test sample x_t in the limit as the inner layer widths grow to infinity. It has been shown that (c.f. (J. Lee et al. 2019; Bietti and Mairal 2019)):

$$f_\infty(\mathrm{x}_t) = f_0(\mathrm{x}_t) - \mathbf{K}_{\mathrm{x}_t\mathbf{X}}\mathbf{K}_{\mathbf{X}\mathbf{X}}^{-1}(f_0(\mathbf{X}) - \mathbf{Y}), \tag{7.3}$$

where $\mathbf{X}$ is the matrix of all training data samples. As before, we consider the case

where the domain is shifted by c. Using (7.3):

$$
\begin{aligned}
f_\infty(\mathrm{x}_t - \mathrm{c}) &= f_0(\mathrm{x}_t - \mathrm{c}) - \mathbf{K}_{(\mathrm{x}_t - \mathrm{c})(\mathbf{X} - \mathrm{c})} \mathbf{K}^{-1}_{(\mathbf{X} - \mathrm{c})(\mathbf{X} - \mathrm{c})} (f_0(\mathbf{X} - \mathrm{c}) - \mathbf{Y}) \\
&\approx f_0(\mathrm{x}_t - \mathrm{c}) - (\mathbf{K}_{\mathrm{x}_t \mathbf{X}} - \Gamma_{\mathrm{x}_t, \mathbf{X}, \mathrm{c}})(\mathbf{K}_{\mathbf{X}\mathbf{X}} - \Gamma_{\mathbf{X}, \mathbf{X}, \mathrm{c}})^{-1} (f_0(\mathbf{X} - \mathrm{c}) - \mathbf{Y}) \\
&= f_0(\mathrm{x}_t - \mathrm{c}) - (\mathbf{K}_{\mathrm{x}_t \mathbf{X}} - \Gamma_{\mathrm{x}_t, \mathbf{X}, \mathrm{c}}) \left(\mathbf{K}^{-1}_{\mathbf{X}\mathbf{X}} + \sum_{m=1}^{\infty} (\mathbf{K}^{-1}_{\mathbf{X}\mathbf{X}} \Gamma_{\mathbf{X}, \mathbf{X}, \mathrm{c}})^m \mathbf{K}^{-1}_{\mathbf{X}\mathbf{X}} \right) (f_0(\mathbf{X} - \mathrm{c}) - \mathbf{Y})
\end{aligned}
\tag{7.4}
$$

$$
\approx \underbrace{f_0(\mathrm{x}_t) - \mathbf{K}_{\mathrm{x}_t \mathbf{X}} \mathbf{K}^{-1}_{X\mathbf{X}} (f_0(\mathbf{X}) - \mathbf{Y})}_{\text{deterministic for fixed } \boldsymbol{\theta}_0} - \underbrace{g(\mathrm{c}, \mathrm{x}_t, \mathbf{X}, \mathbf{Y})}_{\text{random due to c}} \tag{7.5}
$$

In (7.4), we utilize Woodbury's Inverse Identity (Woodbury 1950) to expand the inverse of a sum of matrices. Next, in (7.5), we combine all the terms dependent on c into a function g. Note, we also expand $f_0(\mathrm{x} - \mathrm{c})$ using the Taylor series approximation to represent it as a sum of $f_0(\mathrm{x})$ and other terms.

For a given initialization $\boldsymbol{\theta}_0$, the first term is deterministic, and exactly the same as (7.3), since all other the terms are fixed. However, we see that the second term can vary based on the choice of c. Existing ensembling approaches (Lakshminarayanan, Pritzel, and Blundell 2017) rely on the randomness of the initialization $\boldsymbol{\theta}_0$ to pick diverse solutions from the hypothesis space (He, Lakshminarayanan, and Teh 2020). In contrast, equations (7.2) and (7.5) suggest that even for a fixed $\boldsymbol{\theta}_0$, it is possible to make the NTK stochastic (in c) by shifting the entire input domain. To better understand how the NTK actually changes from these expressions, we study the spectral properties of a simple MLP network empirically, following the analysis in (Tancik et al. 2020).

(b) Spectral properties of shifted NTKs: We compute the Fourier spectra using the same MLP on several shifted domains in Figure 29(B). The original spectrum for the MLP without any shift in the training domain is shown for comparison in 29(A). As indicated by (7.2), we see that each individual shift leads to a different NTK (as

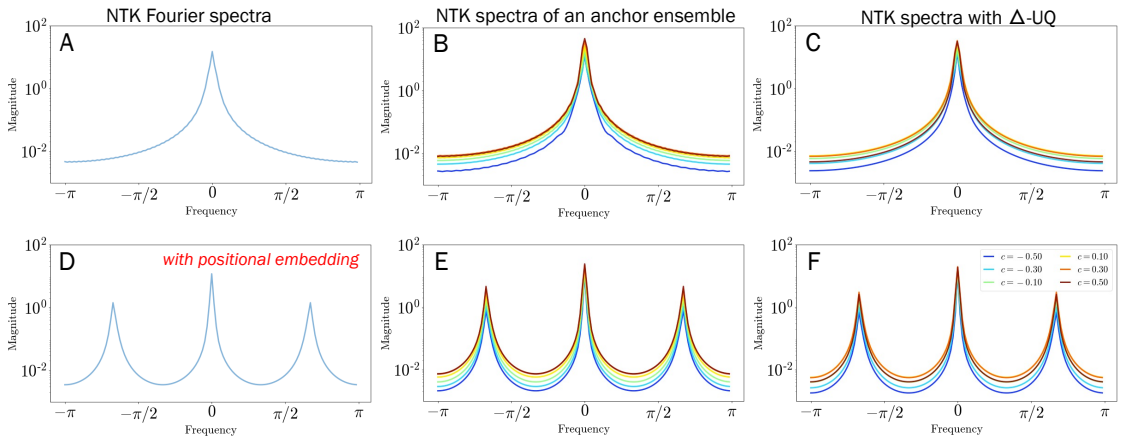


Figure 29. Fourier Spectrum of an NTK for an MLP Model (A,D); Spectra of an Anchor Ensemble (B, E); And NTK Spectra Using $\Delta-$UQ (C, F). Bottom Row Shows NTK Spectra When Inputs Are Passed Through a Sinusoidal PE. We Make Two Key Observations – a) Trivial Shifts in the Input Domain Cause the *Effective* NTK to Be Distinct as a Function of the Shift c, as Seen in Eqn. (7.2); And B) $\Delta-$UQ Achieves a Similar Effect but with a Single Model.

indicated by the spectra), either by flattening it or making it narrower in the frequency domain. The same behavior persists even when we construct positional embeddings (PE), based on sinusoidal functions, prior to building the MLP model (29(D-E)). This leads to one of our main findings stated in 1, and illustrated in Figure 31.

7.4.2 $\Delta-$UQ : Rolling Anchor Ensembles into a Single Model

Since different models in an anchoring-based ensemble are trained with the same initialization, we present a new technique to approximate the epistemic uncertainties using a single neural network. More specifically, we perform a simple coordinate transformation by lifting the domain to a higher dimension as $\mathcal{E} : \mathrm{x} \rightarrow \{\mathrm{c}, \mathrm{x} - \mathrm{c}\}$, we refer to the residual by $\Delta = \mathrm{x} - \mathrm{c}$. This transformation allows the use of multiple representations (w.r.t. different anchors) for the same input sample x, *i.e.*,

Figure 30. Mini-batch Training with $\Delta-$UQ .

```
for inputs, targets in trainloader:
    A = Shuffle(inputs) %% Anchors
    D = inputs—A %% Delta
    X_d = torch.cat([A, D],axis=1)
    y_d = model(X_d) %% prediction
    loss = criterion(y_d,targets)
```

$f_\Delta(\{\mathrm{c}_1, \mathrm{x} - \mathrm{c}_1\}) = f_\Delta(\{\mathrm{c}_2, \mathrm{x} - \mathrm{c}_2\}) = \cdots = f_\Delta(\{\mathrm{c}_k, \mathrm{x} - \mathrm{c}_k\})$, where f_Δ refers to the $\Delta-$UQ model that takes the tuple $(\{\mathrm{c}_k, \mathrm{x} - \mathrm{c}_k\})$ and predicts the target y.

It is easy to see that $[\mathrm{c}, \mathrm{x}_i - \mathrm{c}]^\top [\mathrm{c}, \mathrm{x}_j - \mathrm{c}] = \mathrm{x}_i^\top \mathrm{x}_j - \mathrm{c}^\top(\mathrm{x}_i + \mathrm{x}_j - 2\mathrm{c})$.

That is, the dot product of the transformed inputs takes the same form as before (except for a scaling factor). Therefore, the expressions for the equivalent NTK for different anchor shifts, seen in (7.2), and the corresponding prediction on a test sample seen in (7.5) remain the same for $\Delta-$UQ , by setting $\mathrm{v} = \mathrm{x}_i + \mathrm{x}_j - 2\mathrm{c}$. This leads to our main claim stated in $\boxed{2}$, that the $\Delta-$UQ model achieves similar perturbations of the NTK as an anchor ensemble, based on the choice of c.

Training: During training, for every input x_i we choose an anchor as random sample from the training dataset. Subsequently, we obtain the coordinate transformation $\{[\mathrm{c}, \mathrm{x}_i - c], y_i\}$, using which we train the model. With vector-valued data, this is implemented as a simple concatenation. In the case of image data, we append the channels to create a 6−dimensional tensor (for a 3-channel RGB image). We show simple a Pytorch snippet for training $\Delta-$UQ in Figure 30. Other than increasing the number of parameters in the first layer of the network, $\Delta-$UQ does not incur additional computational overheads.

Over the course of training, every training pair gets combined with a large number

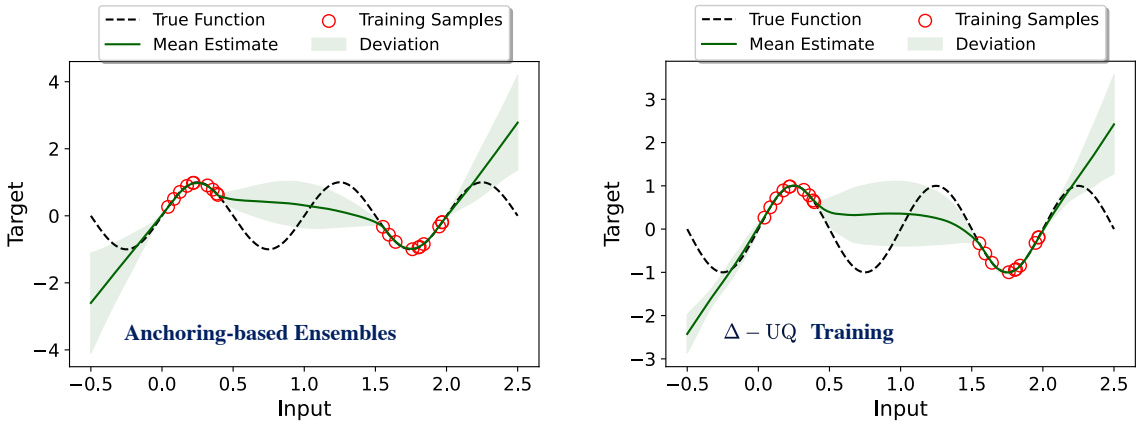


Figure 31. Comparing Anchor Ensembles and $\Delta-$uq in Function Fitting with an MLP. As Expected, We See That the Disagreement Between Models in an Anchor Ensemble Correlate Strongly with the Epistemic Uncertainty, and That $\Delta-$uq , with a Single Model, Matches This Behavior Very Closely.

of anchors. Since the prediction on this training pair – regardless of anchor choice – must always be the same, this places a consistency in the predictions that they must be similar no matter which anchor is chosen. This consistency trades-off with diversity of the kinds of functions that can be learned when compared with an anchor ensemble, where the models are trained independently. This can be seen in the comparisons of the NTK spectra for $\Delta-$UQ with anchor ensembles in Figures 29(C) and (F). In practice, however, we find that the diversity from this single model is still sufficiently large, to estimate good quality uncertainties.

Inference: For a test sample x_t, we obtain the prediction from $\Delta-$UQ as the mean prediction across several randomly chosen anchors; and the standard deviation around these predictions is our estimate for the epistemic uncertainty. In other words, we marginalize out the effect of anchors to obtain the final prediction mean and uncertainty. Formally, the predictive distribution is given by $p(y_t|\mathrm{x}_t) = \int_{\mathrm{c}\in\mathbf{X}} p(y_t|\mathrm{x}_t, \mathrm{c}, \boldsymbol{\theta})p(\mathrm{c})d\mathrm{c}$. In practice, for a trained $\Delta-$UQ model specified as $\boldsymbol{\theta}^*$, we compute the sample mean

and uncertainty around it as:

$$\boldsymbol{\mu}(y_t|\mathrm{x}_t) = \frac{1}{K}\sum_{k=1}^{K} f([\mathrm{c}_k, \mathrm{x}_t - \mathrm{c}_k], \boldsymbol{\theta}^*); \;\; \boldsymbol{\sigma}(y_t|\mathrm{x}_t) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(f([\mathrm{c}_k, \mathrm{x}_t - \mathrm{c}_k], \boldsymbol{\theta}^*) - \boldsymbol{\mu})^2}, \tag{7.6}$$

Discussion: In Figure 31, we show an 1D regression example using 20 training examples along with the predicted mean and estimated uncertainties. As it can be seen, both the anchoring-based ensemble (left) and $\Delta-$UQ training show higher epistemic uncertainties around regions with no training samples. In the ensembles version, we train 20 different networks – while in $\Delta-$UQ we use just a single network, where the uncertainty is obtained using all 20 anchors during inference.

We always only use a single anchor for an input during every training iteration, but multiple anchors during inference time. In theory, multiple anchors could be used during training as well, where the loss is imposed on the mean (obtained with multiple anchors). However, we find that simply using a single random anchor in each iteration achieves a similar effect, as it enforces a consistency that the same training pair (x, y) when combined with many different anchors over the course of training as $[\mathrm{c}_1, \mathrm{x}_i - \mathrm{c}_1], [\mathrm{c}_2, \mathrm{x}_i - \mathrm{c}_2], \ldots, [\mathrm{c}_k, \mathrm{x}_i - \mathrm{c}_k]$ which must all produce the same prediction, y.

$\Delta-$UQ relies on randomly drawn anchors for uncertainty estimation, which is similar to DEns (Lakshminarayanan, Pritzel, and Blundell 2017), that relies on the diversity of the base learners in an ensemble. Arguably, sampling a random set of anchors from the training distribution is simpler than sampling from the posterior $p(\boldsymbol{\theta}|\mathcal{D})$. Furthermore, since every anchor realizes a slightly different function, using a small number of anchors $(5 - 10)$ during inference is typically sufficient to obtain high quality estimates as we show in our experiments. Finally, due to the nature of training with random anchors we also see that $\Delta-$UQ produces particularly effective

uncertainty estimates when the training set size is small, and this proves to be very useful in applications like sequential optimization.

7.5 Experiments and Findings

We validate our approach in this section using a variety of applications and benchmarks – (a) first, we consider the utility of epistemic uncertainties in object recognition problems where they have been successfully used for outlier rejection and calibrating models under distribution shifts. We show that $\Delta-$UQ can be very effective even with large-scale datasets like ImageNet (Russakovsky et al. 2015); (b) next, we consider the challenging problem of sequential design optimization of black-box functions, where the goal is to maximize a scalar function of interest with the fewest number of sample evaluations. Using a Bayesian optimization setup, we show that the uncertainties obtained using $\Delta-$UQ outperform many competitive methods across an extensive suite of black-box functions.

7.5.1 Outlier Rejection

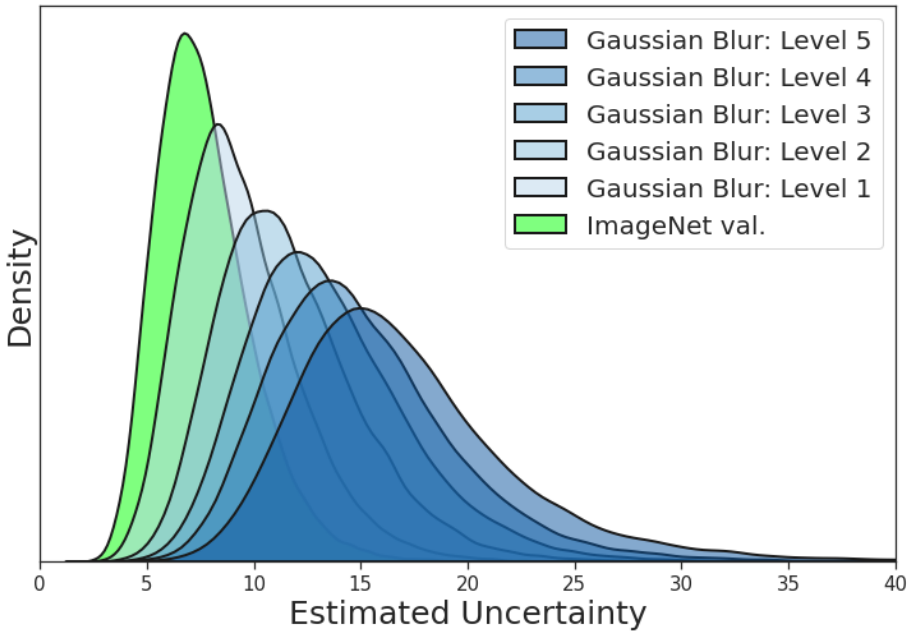
A popular application for epistemic uncertainties is in rejecting outliers since, by definition, they are in regions outside of the training distribution. As such, we expect the model to produce highly uncertain predictions for these images, which should help us design an effective OOD detector. To evaluate this hypothesis, we train a modified ResNet-50 (He et al. 2016) model on ImageNet that accepts 6 input channels (anchor, Δ) as outlined earlier. We train the model using standard hyperparameter settings except, training it longer for 120 epochs – our top-1 accuracy (76.1) matches that of a

Table 8. Calibration under Distribution Shift:- a Resnet-50 Model That Is Tempered by Uncertainties Obtained from $\Delta-$UQ (See Text) Outperforms Several Competitive Baselines Averaged Across 16 Different Corruptions of Imagenet-C at Highest Severity Level 5.

Metric		Vanilla	Temp Scaling	DEns	MCD	LL-Dropout	SVI	SVI-AvUC	Ours
ECE ↓	lower quartile	0.124	0.096	0.050	0.078	0.093	0.072	0.032	0.022
	median	0.174	0.139	0.090	0.134	0.145	0.114	0.045	0.038
	mean	0.194	0.160	0.088	0.153	0.161	0.119	0.054	0.044
	upper quartile	0.274	0.236	0.126	0.219	0.236	0.172	0.070	0.063
NLL ↓	lower quartile	4.635	4.53	4.035	4.699	4.563	4.322	4.164	4.014
	median	5.115	4.993	4.624	5.093	5.034	4.853	4.823	4.617
	mean	5.234	5.091	4.604	5.553	5.201	4.865	4.707	4.352
	upper quartile	6.292	6.165	5.893	6.522	6.342	6.034	5.778	4.987
Brier ↓	lower quartile	0.941	0.926	0.877	0.933	0.923	0.906	0.883	0.868
	median	0.987	0.970	0.922	0.967	0.969	0.943	0.935	0.925
	mean	0.964	0.945	0.888	0.961	0.947	0.922	0.900	0.887
	upper quartile	1.052	1.027	0.989	1.025	1.025	1.013	0.985	0.949

standard ResNet-50. Specifically for images, we found that corrupting the anchors with common transforms like random crops, Gaussian blurs, and color jitter improves performance. That is, instead of $[c, x - c]$, we use $[\mathcal{T}(c), x - c]$ where $\mathcal{T}$ is a transform like Gaussian blur. We describe this process in more detail in the supplement. The uncertainty for a test sample is given by the standard deviation of the logits obtained by varying the anchors. To obtain a scalar statistic, we compute the mean across all classes as the uncertainty score for that sample.

We show the results for outlier rejection in Table 32b, where we follow the protocol



(a) Uncertainties Change Meaningfully as Outliers Get More Severe

Method	AUROC ↑	DTACC ↑	AUPR-in/out ↑
ResNet-50 (He et al. 2016)	93.36	86.08	92.82 / 93.71
Temp-Scal (Guo et al. 2017)	93.71	86.47	93.21 / 94.01
Deep Ens (Lakshminarayanan, Pritzel, and Blundell 2017)	95.49	88.82	95.31 / 95.64
MC Dropout (Gal and Ghahramani 2016)	96.38	89.98	96.16 / 96.67
SVI (Blundell et al. 2015)	96.40	90.03	95.97 / 96.83
$\Delta-$UQ (ours)	**97.49**	**91.90**	**97.56** / **97.47**

(b) Uncertainties from $\Delta-$UQ for Outlier Rejection

Figure 32. Rejecting Outliers with Epistemic Uncertainties:- We Evaluate $\Delta-$UQ on the Benchmark Introduced by (Krishnan and Tickoo 2020) Where We Use Gaussian Blur of Level 5 Intensity as the Outliers from the Imagenet Validation Set. At Inference, Uncertainties Are Estimated as the Mean of Std. Dev of Predictions Obtained with 10 Anchors.

established in (Krishnan and Tickoo 2020), that uses a Gaussian blur of intensity 5 from ImageNet-C (Hendrycks and Dietterich 2019) as the outlier set, and the clean ImageNet validation data as inliers. We use the estimated uncertainty obtained with

10 anchors as in (7.6) as our score for outlier rejection and report commonly used metrics such as AUROC, Detection Accuracy (DTACC), and AUPR-in/out. We note that, just the inconsistency of predictions obtained using $\Delta-$UQ outperforms many baselines including mean-field stochastic variational inference (SVI) (Graves 2011; Blundell et al. 2015), SVI-AvUC (Krishnan and Tickoo 2020), Monte Carlo dropout (Gal and Ghahramani 2016), and temp. scaling (Guo et al. 2017). While this can be further improved by taking the mean prediction into account, similar to existing approaches for semantic novelty detection (with scores such as entropy, energy (Liu et al. 2020) etc.), our focus here is to evaluate the quality of uncertainty alone. In Figure 32a, we observe that $\Delta-$UQ 's uncertainty estimate changes smoothly as the outliers become farther away (more severe intensity) from the training distribution.

7.5.2 Calibration under Distribution Shifts

Following our observation that our uncertainty estimates are effective in rejecting outliers in table 32b, here we study if they can be leveraged to calibrate ImageNet models under distribution shift. To calibrate a classifier, we simply scale the logits of the mean by the uncertainties as follows: $\boldsymbol{\mu}_{\text{calib.}} = \boldsymbol{\mu}(1 - \bar{\boldsymbol{\sigma}})$, where $\bar{\boldsymbol{\sigma}}$ is the standard deviation estimated from (7.6), where once again for classification we simply compute the average of standard deviation across all the 1000 classes, followed by min-max normalization to $[0, 1)$. This simple scaling of the mean reflects our prior belief – a highly certain prediction must remain unchanged, whereas an uncertain one gets tempered down. Note, this scaling is applied to logits from all classes and hence the accuracy of the mean remains unchanged before or after calibration. We evaluate how calibrated the predictions are using three commonly used metrics – Calibration error

(ECE), negative log likelihood (NLL), and Brier Score. We use the same ResNet-50 classifier trained on ImageNet as before, and measure calibration for predictions on 16 different ImageNet-C corruptions at severity 5. We list the corruptions and show examples of them in the supplement. We report the 25^{th} (lower quartile), 50^{th} (median) and 75^{th} (upper) quantiles along with the mean of the three metrics across 16 corruptions in table 8. Across all measures we see that $\Delta-\text{UQ}$ is able to calibrate models better, even in comparison to state-of-the-art approaches that use an explicit calibration objective to adjust the prediction probabilities, further validating the quality of its uncertainty estimates.

7.5.3 Sequential Optimization

Denoting a high-dimensional function as $\text{f} : \mathcal{D} \rightarrow \mathbb{R}$, our goal is to solve the following optimization problem: $\text{x}^* = arg\max_{\text{x}\in\mathcal{D}} \text{f(x)}$. Here, $\mathcal{D}$ refers to a *bounded* design space comprising D different parameters with their corresponding value ranges $[\ell_d, h_d], \forall d = 1 \cdots D$. The high computational or financial cost of evaluating f (invoking a simulator or running an experiment) motivates the additional objective of minimizing the number of evaluations.

Given the high-dimensional nature of the design spaces, a simple brute-force search or even space-filling random sample designs (Kailkhura et al. 2018) often require significantly large sample sizes to identify the optima, thus motivating the use of *sequential optimization* strategies. In particular, Bayesian Optimization (BO) techniques based on statistical surrogates (e.g., Gaussian processes) form an important class of solutions (Shahriari et al. 2015). In a nutshell, given an initial experiment design and their function evaluations, $\{\text{x}_i, f(\text{x}_i)\}_{i=0}^{n_0}$, sequential optimization techniques

Table 9. Sequential Optimization:- We Rigorously Evaluate the Performance of Different Uncertainty Estimators on a Suite of Black-box Functions and Report the AUC Metric (↑) Averaged Across Multiple Random Seeds and Trials. In Each Case, We Also Indicate the Number of Initial Samples and Optimization Steps.

Function	Dim.	Init.	Steps	GP	MCD	BNN	DEns	Ours
Multi Optima	1	5	25	0.51 ± 0.2	0.45 ± 0.16	0.64 ± 0.12	0.28 ± 0.17	0.73 ± 0.09
Ackley	2	5	25	0.23 ± 0.08	0.76 ± 0.03	0.71 ± 0.1	0.75 ± 0.04	0.83 ± 0.03
Beale	2	5	25	0.64 ± 0.31	0.55 ± 0.22	0.27 ± 0.17	0.81 ± 0.03	0.85 ± 0.04
Booth	2	5	25	0.39 ± 0.21	0.55 ± 0.14	0.3 ± 0.2	0.68 ± 0.06	0.79 ± 0.04
Branin	2	5	25	0.35 ± 0.28	0.28 ± 0.19	0.22 ± 0.14	0.46 ± 0.1	0.67 ± 0.06
Bukin	2	5	25	0.36 ± 0.12	0.55 ± 0.07	0.38 ± 0.11	0.59 ± 0.11	0.76 ± 0.1
Camel	2	5	25	0.83 ± 0.08	0.86 ± 0.06	0.84 ± 0.03	0.83 ± 0.07	0.89 ± 0.03
Dropwave	2	5	25	0.68 ± 0.15	0.57 ± 0.18	0.67 ± 0.13	0.67 ± 0.11	0.79 ± 0.14
Griewank	2	5	25	0.83 ± 0.02	0.74 ± 0.04	0.59 ± 0.17	0.7 ± 0.14	0.86 ± 0.03
Holder	2	5	25	0.12 ± 0.06	0.36 ± 0.28	0.36 ± 0.37	0.39 ± 0.29	0.57 ± 0.07
Levi N.13	2	5	25	0.26 ± 0.26	0.75 ± 0.1	0.7 ± 0.1	0.6 ± 0.11	0.87 ± 0.07
Levy	2	5	25	0.57 ± 0.18	0.61 ± 0.32	0.55 ± 0.16	0.59 ± 0.16	0.83 ± 0.03
Hartmann	3	5	25	0.57 ± 0.07	0.49 ± 0.11	0.46 ± 0.16	0.53 ± 0.15	0.68 ± 0.07
Ackley	4	10	25	0.17 ± 0.02	0.06 ± 0.04	0.1 ± 0.03	0.14 ± 0.02	0.59 ± 0.05
Griewank	4	10	25	0.37 ± 0.08	0.47 ± 0.06	0.39 ± 0.05	0.43 ± 0.07	0.69 ± 0.07
Levy	4	10	25	0.1 ± 0.08	0.4 ± 0.3	0.27 ± 0.21	0.21 ± 0.1	0.62 ± 0.2
Hartmann	6	10	25	0.15 ± 0.01	0.2 ± 0.08	0.1 ± 0.05	0.15 ± 0.04	0.27 ± 0.15
Ackley	8	10	50	0.06 ± 0.09	0.09 ± 0.13	0.08 ± 0.12	0.11 ± 0.05	0.36 ± 0.09
Griewank	8	10	50	0.07 ± 0.02	0.12 ± 0.13	0.08 ± 0.07	0.19 ± 0.07	0.32 ± 0.11
Levy	8	10	50	0.12 ± 0.04	0.16 ± 0.08	0.11 ± 0.07	0.13 ± 0.07	0.47 ± 0.03
Avg. Rank	-	-	-	3.8	3.0	4.1	2.95	1.0

incrementally select candidates to achieve the so-called *exploration-exploitation* trade-off using an appropriate *acquisition* function (Snoek, Larochelle, and Adams 2012). In this study, we use the popular expected improvement (EI) score to perform candidate selection.

Setup: In this experiment, we consider a large suite of black-box optimization functions with varying dimensionality (1 to 8) and complexity to comprehensively evaluate

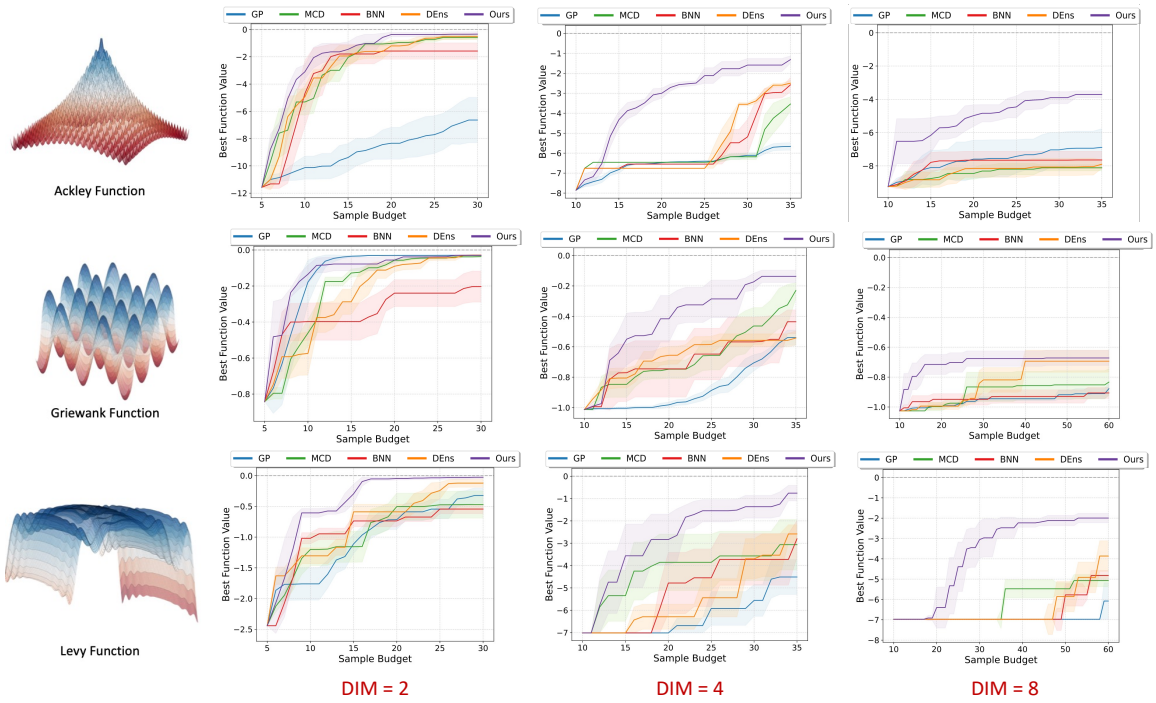


Figure 33. Convergence Curves Obtained with Different Uncertainty Estimation Methods:- We Show the Best Function Value Achieved for Three Different Functions at Dimensions 2, 4 and 8 Respectively (for 1 Random Seed, 5 Trials). We Find That $\Delta-$UQ Consistently Outperforms All Other Baselines. The Effectiveness of Our Approach in Producing Meaningful Uncertainties at Small Sample Sizes Becomes More Apparent as Dimensionality Increases.

different epistemic uncertainty estimation techniques with deep neural network surrogates (see Appendix 1 for description of the functions used in our study). Finally, we perform an experiment with a pre-trained generative model (GAN) trained on MNIST handwritten digits, wherein we perform optimization in the 100$-$D latent space $\mathcal{Z}$ such that thickness of the resulting digit is maximized: $\sum_i \mathbb{I}(x_i > 0), \forall i$, where $\mathbb{I}$ denotes the identity function.

We use the following baseline uncertainty estimation approaches in our study: (i) Gaussian processes (GP); (ii) Monte-Carlo dropout (MCD); (iii) Bayesian neural networks (BNN) trained via variational inferencing; and (iv) deep ensembles (DEns).

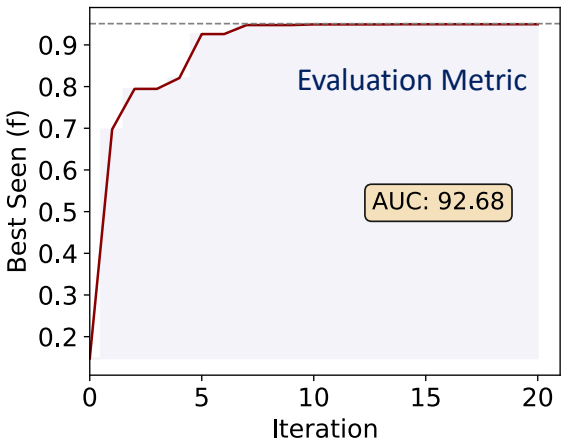
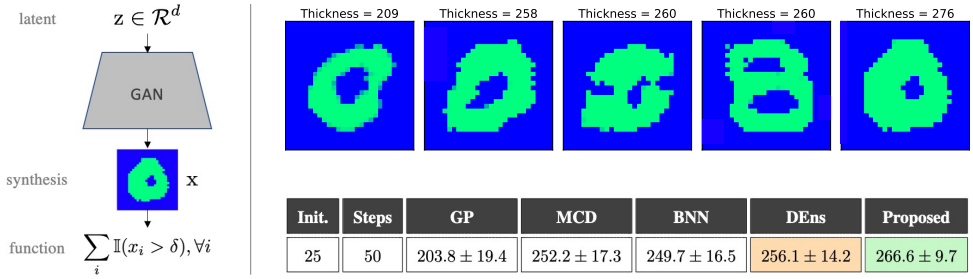


Figure 34. Area under the Curve (AUC) Metric for Evaluating Sequential Optimization Performance.

For all neural network surrogates, we computed positional embeddings (sinusoidal) of the raw parameter inputs prior to building a fully-connected network with 4 hidden layers each containing 128 neurons and ReLU activation. All methods were trained with the same set of hyperparameters: Adam optimizer learning rate $1e-4$ and 500 epochs, except for BNN, which required 1000 epochs for convergence. With MCD, we used 50 forward passes at test time for each sample to obtain the uncertainties. Finally, with $\Delta-$UQ, we set the number of anchors for inferencing as $\min(20, n)$, where n is the number of samples in the observed dataset in any iteration.

The DEns model was constructed using 5 constituent members (increasing this did not provide any benefits), each trained with a different initialization. The number of initial samples and the number of steps in the sequential optimization were set to be the same across all methods. In each round of the Bayesian optimization, we used $10,000$ samples for initialization and 15 restarts (i.e., starting points for multistart acquisition function optimization), and finally one candidate ($q = 1$) was evaluated with the black-box function and added to the observed dataset. We performed experiments with 5 random seeds (different initializations), each for 5 independent trials. Since



Init.	Steps	GP	MCD	BNN	DEns	Proposed
25	50	203.8 ± 19.4	252.2 ± 17.3	249.7 ± 16.5	256.1 ± 14.2	266.6 ± 9.7

Figure 35. GAN-based Optimization:- $\Delta-$UQ Consistently Produces Images with Higher Function Values (Thickness) for the Same Sampling Budget, When Compared to Existing Baseline Methods.

the goal is to reach the global optima with the fewest number of samples, we use the widely adopted area under the *iteration vs best achieved function value* curve to obtain a holistic evaluation of different approaches (see Figure 34).

Results: From table 9, we find that $\Delta-$UQ produces significantly higher AUC scores in comparison to existing baselines, across all benchmark functions. While MCD and DEns behave reasonably well in low dimensions, their performance suffers when we go to higher dimensions (see figure 33). Furthermore, we find that the performance of BNN is generally lower due to the inherent small samples sizes that we operate in. Finally, as showed in figure 35, our approach consistently achieves higher function values in the MNIST GAN-based optimization, thus validating the quality of the uncertainties produced via anchoring.

7.6 Summary

In this chapter, we identified that an ensemble trained on datasets with trivial constant bias gives rise to slightly inconsistent trained models. By utilizing such a principle we designed $\Delta-$UQ, a simple, scalable, and accurate single model uncertainty

estimator that outperforms many existing techniques. We demonstrated the efficacy of our methods on a wide variety of benchmarks that rely on accurate epistemic uncertainty estimation such as sequential optimization and outlier rejection.

Chapter 8

PREDICTING GENERALIZATION GAP IN DEEP MODELS VIA REPRESENTATION UNCERTAINTIES FROM Δ-UQ

While it is critical to effectively estimate predictive uncertainties to provide useful signals for model confidence, producing signals on shifted data regimes where a model can fail to generalize is crucial in understanding how DNNs behave in uncontrollable 'in-the-wild' scenarios. For example, a DNN trained on a given source dataset can offer varying levels of generalization for data with large distribution shifts from the source. It therefore becomes imperative to provide an accurate estimation of expected generalization accuracy when such pre-trained DNNs are utilized on completely unlabeled, 'in-the-wild' target distributions. In this chapter, we propose a novel strategy for directly predicting accuracy on unseen target data with the help of anchoring in Δ-UQ introduced in Chapter 6. Anchoring has been shown previously to perform effectively in characterizing domain shifts, which we exploit for estimating data representation uncertainties as proxies for the generalization gap. We systematically show that Δ-UQ can effectively capture reliable uncertainties even under challenging distribution shifts.

8.1 Problem Setup

With tremendous success exhibited by AI methods (Dosovitskiy et al. 2021; Devlin et al. 2019; Vaswani et al. 2017), off-the shelf black-box neural networks are being increasingly deployed to guide decision making even in critical applications such as

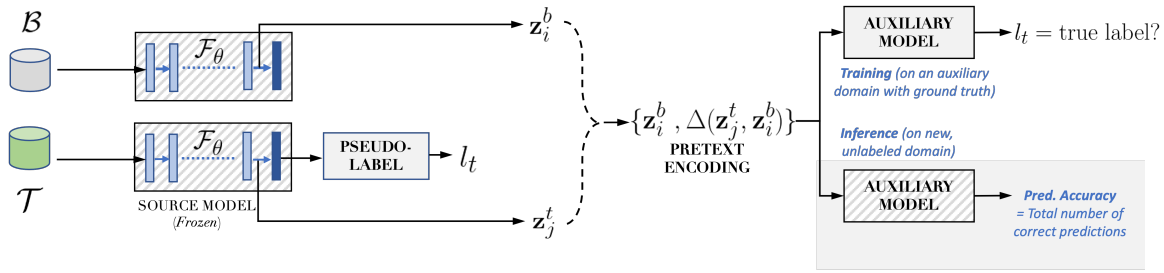


Figure 36. Overview of Our Approach to Predict Accuracy on Unseen Target Distributions. Utilizing the Intermediate Features $\mathbf{z}_i^b$, $\mathbf{z}_j^t$ Extracted from Samples of the Source $\mathcal{B}$ and Target $\mathcal{T}$ Distributions Respectively from a Pre-trained Classifier $\mathcal{F}$, We Construct Pre-text Encodings of the Form $\{\mathbf{z}_i^b, \Delta(\mathbf{z}_j^t, \mathbf{z}_i^b)\}$ to Train Auxiliary Models That Can Be Used to Predict Generalization Accuracy. This Encoding Strategy Effectively Captures the Important Differences Between the Source and Target Distributions Which Can Be Leveraged to Estimate Generalization Gaps.

healthcare. However, these models have been shown to function reliably only when the test distribution overlaps significantly with the training distribution (Ben-David et al. 2010; Recht et al. 2019) which is seldom the case, making it difficult to safely rely on such model predictions on new test domains. An important step towards promoting the adoption of these models in practice is not only to ensure that they behave predictably on regimes where the training data provides meaningful evidence, but also to provide an accurate estimation of expected generalization accuracy when utilized on completely unlabeled, 'in-the-wild' target distributions.

Estimating generalization accuracy on target distributions is an important topic of research (Jiang et al. 2019; Guillory et al. 2021; Deng and Zheng 2021). For instance, Jiang *et al.* (Jiang et al. 2019), showed that distances between the training distribution and model decision boundaries can be a strong indicator to predict generalization error for unseen examples. However, this approach makes the assumption that the train and test distributions are similar, and hence the predicted generalization gap need not be accurate under distribution shifts. Recently developed strategies include training a post-hoc predictor based on metrics used commonly in domain adaptation

settings (Gretton et al. 2012; Glorot, Bordes, and Bengio 2011) that quantify differences in data distributions to predict the generalization performance. For e.g., Deng *et al.* (Deng and Zheng 2021) use the Frechét distances between the training and the target set to fit regression models to estimate accuracy. The authors demonstrate the existence of linear relationships between accuracy gap and the distribution distances. On similar lines, Guillory *et al.* (Guillory et al. 2021) utilize the difference of confidence (DoC) metric between the train and the target sets in order to predict change in accuracies on unseen datasets and show that DoC significantly outperforms other distribution difference metrics over a variety of natural and synthetic distribution shifts. Despite the simplicity of these methods, we find that there still exists a significantly wide accuracy gap when applied on real-world domain shifts as the regressor is not guaranteed to be calibrated well enough to reflect the uncertainties between the train and target domains.

Uncertainty estimation in machine learning (Thiagarajan, Venkatesh, Sattigeri, et al. 2020) is a powerful tool to characterize model behaviour under distribution shifts and identify regimes of improper sampling in data. We demonstrated $\Delta-$UQ in Chapter 6 as an accurate, efficient single model uncertainty estimator. $\Delta-$UQ is based on *anchoring* — where the input is transformed into a tuple consisting of an anchor sample (random sample drawn from a prior) and a pretext encoding of the input with the anchor to train predictive models. During inference, an anchor marginalization strategy is used to obtain the prediction for each sample over multiple randomly chosen anchors. We showed that, $\Delta-$UQ is a powerful mechanism to effectively distinguish between in and out-of-distribution samples, in contrast to stochastic approaches such as deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017). Interestingly, we observe that anchoring a predictive model with the pre-text task as done in Chapter

6, implicitly defines a metric between two distributions – the anchor distribution and the source distribution on which the model is trained. As a result, we hypothesize that an such a strategy can in fact provide meaningful information (representation uncertainty) regarding shifts between different data distributions and can be leveraged in order to directly predict the generalization accuracy on unseen target distributions – without relying on summary metrics.

Our Work: In this work, we propose a novel strategy for directly predicting accuracy of deep models on previously unseen, unlabeled target distributions. Specifically, we utilize the $\Delta-$encoding scheme from (Anirudh 2021) to train an auxiliary model, which is comprised of two main components – (a) a decoder model that tries to undo the pretext encoding to recover the representation of a target sample obtained from the pre-trained model, implicitly capturing the relationship between the source and target datasets; and (b) a binary classifier that acts on the residual between the decoded sample and the target representation, to predict if the source network correctly classified this sample or not. By training this entire process end-to-end (while keeping the source model frozen), our experiments on synthetic and the multi-domain PACS dataset (Li et al. 2017) show that, in addition to providing well-calibrated target accuracy estimates, our approach also outperforms existing baselines.

8.2 Approach

In this section, we describe our approach for predicting accuracy of a trained model on unseen target distributions. An overview of our approach is illustrated in Figure 36.

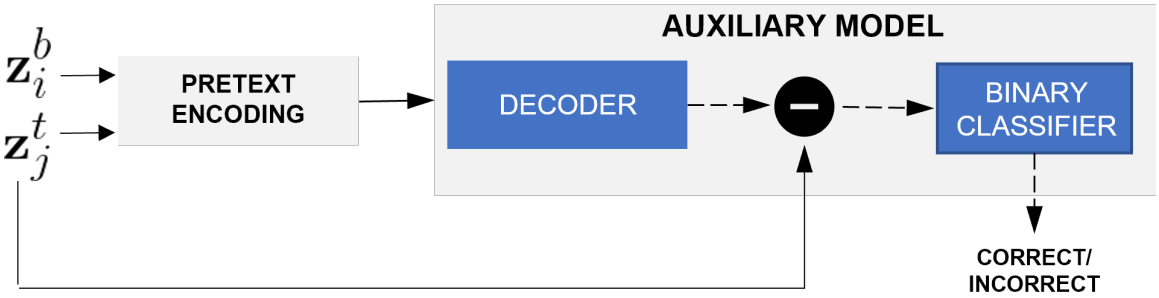


Figure 37. The Auxiliary Model Block Consists of Two Components (i) a Decoder That Tries to Undo the Pre-text Encoding to Recover a Representation of the Target Sample Obtained from the Pre-trained Model Capturing the Relationships Between the Source and Targets (ii) a Binary Classifier to Predict Whether the Target Sample Has Been Correctly Classified or Not by $\mathcal{F}$.

8.2.1 Preliminaries and Notations

Let $\mathcal{F}_\theta$ denote a multi-class classifier (source model) parameterized by θ that takes an image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ as input to predict the output label $\hat{\mathrm{y}} \in \mathcal{Y} \coloneqq [1, \cdots, K]$. Here, $\mathbf{x}$ is a C channel image of height H and width W and K represents the total number of classes. Let $\mathcal{D}_\Phi$ denote the decoder network with parameters Φ and $\mathcal{G}_\Psi$ corresponds to the binary classifier with parameters Ψ. In this paper, we assume that the $\mathcal{F}_\theta$ is pre-trained on the source distribution $\mathcal{B}$ consisting of N samples $\{x_i^b,\ y_i^b\ |i = 1 \dots N\} \sim \mathcal{B}$. Let $\mathcal{T}$ denote the target distribution consisting of M samples $\{x_j^t,\ y_j^t\ |j = 1 \dots M\} \sim \mathcal{T}$. The set of labels present in both the base and target distributions are assumed to be identical. Additionally, we define an additional calibration dataset, $\mathcal{T}_c \neq \mathcal{B}$, which is used to train $\mathcal{D}_\Phi$ and $\mathcal{G}_\Psi$. At inference time, samples from the target domain are passed through the trained models to finally obtain an estimate of accuracy of $\mathcal{F}_\theta$. Note, in all of this the pre-trained model $\mathcal{F}_\theta$ is kept frozen.

8.2.2 Pretext Encoding Scheme

We define pretext encoding as a function $\Delta(Q, R)$, where Q denotes a query sample and R is an anchor with the same dimensions as Q, randomly drawn from a distribution $P(R)$. We reformulate the problem of training the auxiliary models using tuples constructed using anchors and the pre-text encodings $\{R, \Delta(Q, R)\}$. The tuple is realized by concatenating R and $\Delta(Q, R)$ (in the channel dimension). Following (Anirudh 2021), we consider the pretext encoding function as $\Delta(Q, R) = Q - R$.

In our approach, we construct the tuples using latent features computed for the base and target distributions using $\mathcal{F}_\theta$, *i.e.*, $R = \mathbf{z}_i^b$ and $Q = \mathbf{z}_j^t$. For training the models, we randomly choose a single anchor R for every input sample Q in a batch to ensure that each input sample is combined with different random anchors as the training progresses.

8.2.3 Training the Auxiliary Models

8.2.3.1 Intuition

The auxiliary models which include the decoder $\mathcal{D}_\Phi(.)$ and the binary classifier $\mathcal{G}_\Psi(.)$ are critical to evaluate generalization performance on unseen target data (Figure 37). The tuples extracted from the pretext encoding stage are used to jointly train $\mathcal{D}_\Phi$ and $\mathcal{G}_\Psi$. The key intuition behind our methodology is based on the idea that when the target representation is encoded with randomly chosen source representations during training, inconsistencies in recovering the target will induce large representation uncertainties which manifest as the appropriate prediction by the binary classifier.

Based on such uncertainties which are richer metrics than summary statistics adopted in state-of-the-art approaches (Guillory et al. 2021) for characterizing domain shifts, the binary classifier predicts whether the target has been correctly identified by the pre-trained classifier or not.

In our construction, the decoder is stochastic (input Q can be associated with any random R during training). The representation $\mathbf{d}$ from the decoder can be interpreted as a 'reconstruction' of the input Q by averaging over different choices of R. When $\mathbf{d}$ exactly matches Q there exists no uncertainty in the input. On the other hand, if there is a mismatch, it denotes the epistemic uncertainty in recovering Q over the distribution of anchors.

By selecting anchors from the base distribution $\mathcal{B}$ and query samples from the target distribution $\mathcal{T}$, wherein the two datasets can have non-overlapping manifolds, recovering the input samples under different anchor choices can lead to larger discrepancies which can be exploited to detect domain shifts.

8.2.3.2 Decoder

In our work, the decoder $\mathcal{D}_\Phi$ operates on the tuple to produce an intermediate representation $\mathbf{d}$ with the same dimensions as that of $\mathbf{z}_j^t$. We then compute residuals $|\mathbf{d} - \mathbf{z}_j^t|$ (indicative of the uncertainties) to train the binary classifier $\mathcal{G}_\Psi$.

8.2.3.3 Binary Classifier

In order to train the binary classifier, which can be eventually used to estimate the generalization performance of the underlying pre-trained classifier, we obtain labels

from $\mathcal{F}_\theta$ indicative of whether the data sample x_j^t has been correctly classified or not. We first determine the true class likelihoods $P(y_j^t|x_j^t)$ and perform the following pseudo-labeling strategy to prepare the labels for the binary classifier.

$$l_t = \begin{cases} 1, & \text{if } P(y_j^t|x_j^t) >= \tau_2, \\ 0, & \text{if } P(y_j^t|x_j^t) <= \tau_1, \\ \text{ignore}, & \text{otherwise.} \end{cases} \tag{8.1}$$

Here, τ_1, τ_2 are user-specified thresholds. The binary classifier which then estimates the likelihoods $\hat{p}_j^t$ of the target data samples being correctly classified by $\mathcal{F}_\theta$ is jointly trained with the decoder using the Binary Cross Entropy loss function.

8.2.4 Predicting Generalization

During inference with the pre-trained classifier $\mathcal{F}$ and auxiliary models $\mathcal{D}$ and $\mathcal{G}$, we estimate the mean likelihood of the unseen target data sample being correctly classified by analyzing the consistency in the prediction with respect to K different anchors drawn from the base distribution $\mathcal{B}$. Therefore for a given target sample, we marginalize the impact of the anchors and compute the prediction as follows:

$$\overline{\widehat{p}_j^t} = \frac{1}{K} \sum_K \mathcal{G}(|\mathcal{D}(\mathbf{z}_k^b, \Delta(\mathbf{z}_j^t, \mathbf{z}_k^b)) - \mathbf{z}_j^t|), \tag{8.2}$$

We finally estimate the overall generalization accuracy on the entire unseen target dataset, by aggregating the individual mean predictions as follows:

$$\text{Acc}(x_j^t|j = 1 \dots M) = \frac{1}{M} \sum_j \mathbb{I}(\overline{\widehat{p}_j^t} > \gamma), \tag{8.3}$$

where γ is the threshold of detection.

8.3 Experiment Setup

8.3.1 Datasets

We evaluate our approach for predicting generalization performance on unseen target domains and synthetic variations in the Photo-Art-Cartoon-Sketch (PACS) (Li et al. 2017) dataset. The dataset contains 9991 images across the different domains. All domains share the same label space where each image is associated with one of the following 7 categories namely person, house, horse, guitar, giraffe, elephant and dogs. In all our experiments we utilize the images from the photo domain as the base distribution $\mathcal{B}$ and the images from the sketch domain as the auxiliary domain $\mathcal{T}_c$ to train the auxiliary models. The cartoon and the art domains are considered only for predicting the generalization performance. Further, we perform synthetic augmentations onto the photo, cartoon, and the art domains to create datasets with different distribution shifts such horizontal flips, brightness and hue, random rotations and Gaussian blur of low severity. In total, we evaluate our model, and baselines on a total of 14 test domains containing both natural and synthetic distribution shifts, to predict the generalization gap.

8.3.2 Setup

We use a Resnet-18 model (He et al. 2016) pre-trained on Imagenet (Deng et al. 2009) as the classifier $\mathcal{F}$ and extract the latent features from the last residual block and perform average pooling to obtain $\mathbf{z}_i$. The decoder $\mathcal{D}$ and the binary classifier $\mathcal{G}$ are fully connected neural networks with 5 hidden layers of 512 neurons and 2 hidden

layers of 512 neurons respectively. All models are trained using the ADAM optimizer with a learning rate of $3e^{-4}$ for training. We use predictions obtained with the source model $\mathcal{F}$ as pseudo-labels to train the auxiliary model, and use thresholds $\tau_1 = 0.25$ and $\tau_2 = 0.65$ in an attempt to better guide the auxiliary model training process, as explained in (8.1). In order to obtain the final prediction from the binary classifier, we use a threshold, $\gamma = 0.6$, which we found to be sufficiently conservative in all our experiments.

8.3.3 Baselines

(i) *Difference of Confidence (DoC)* (Guillory et al. 2021). DoC is a recently proposed approach for predicting generalization gaps on unseen distributions that has been shown to outperform existing baselines that operate on conventional distributional distances such as Maximum Mean Discrepancy (MMD) (Gretton et al. 2012) and Frećhet distances. Following (Guillory et al. 2021), we fit a linear regressor to the DoC scores to predict change in accuracies by constructing random subsets of the source domain (Photo) and the calibration domain (Sketch) with different synthetic augmentations and evaluate the regressor on all unseen distributions. (ii) *Auxiliary model training w/o pretext encoding and anchoring.* As an ablation, we consider a baseline that directly takes the intermediate feature representation for the auxiliary domain, and predicts whether or not the main ResNet model, $\mathcal{F}$, got the prediction correct. We find that this model tends to overfit easily, so we simply use the binary classifier portion of our main model as the auxiliary model.

Table 10. Means and Standard Deviation of the Accuracy Gaps over Different Target Distributions.

Methods	**|True - Predicted Accuracy| Mean ± Std**
DOC	0.3650 ± 0.2903
w/o Δ-Encoding	0.211 ± 0.113
Ours	**0.12 ± 0.06**

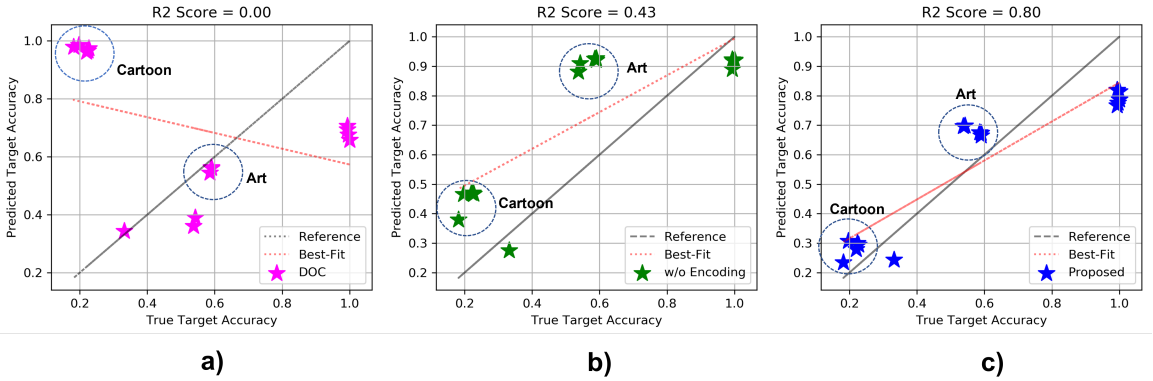


Figure 38. Comparison of the Generalization Performance of Our Approach on Different Target Domains and Synthetic Shifts over Existing Baselines. We Find That Our Approach Reliably Estimates the Generalization Accuracy with a Strong Linear Relation Between the True and Predicted Target Accuracies Against the Baseline Approaches.

8.4 Results and Discussion

In Fig. 38, we show true and predicted accuracies for 16 test domains (14 unseen, and test splits of photo and sketch used as training and auxiliary domains respectively) across the three techniques considered here. We find that our method significantly outperforms the others significantly, with a strong linear relationship between the true and predicted accuracies, resulting in an R2 score of 0.8. This is in comparison to 0.43 for the ablation model without pretext encoding. Interestingly, we also note that the ablation model also suffers from mostly being over confident in its predictions (most of the values are above the diagonal), whereas the anchoring and pretext encoding is

less biased in its estimates on average. The DoC (Guillory et al. 2021) predictions shown in Figure 38 (left most) suffer primarily due to its failure on domains that are significantly shifted from the source or auxiliary domains. Corroborating Figure 38, Table 10, provides the mean and standard deviations of the accuracy gaps over different target distributions. It can be observed that our approach outperforms the baselines in producing significantly low accuracy gaps.

8.5 Summary

In this chapter, we developed a novel strategy for directly predicting accuracy of a deep neural network on unseen target distributions based on Δ-UQ. We find anchoring to be an effective strategy to train post-hoc accuracy estimators, that are able to better model the distribution shifts as a function of the generalization gap, unlike existing methods that rely on difference metrics. In particular, we train auxiliary models containing (a) decoder that tries to undo the pretext encoding to recover the representation of a target sample and (b) a binary classifier that acts on the residual between the decoded sample and the target representation, to predict if the source network correctly classified this sample or not. Through extensive experiments on the PACS dataset along with synthetic variations, we found that our approach provided reliable accuracy estimates outperforming the existing baseline and ablations.

Chapter 9

UNSUPERVISED AUDIO SOURCE SEPARATION WITH SOURCE SPECIFIC GENERATIVE PRIORS

State-of-the-art under-determined audio source separation systems rely on supervised end-end training of carefully tailored neural network architectures operating either in the time or the spectral domain. However, these methods are severely challenged in terms of requiring access to expensive source level labeled data and being specific to a given set of sources and the mixing process, which demands complete re-training when those assumptions change. This strongly emphasizes the need for unsupervised methods that can leverage the recent advances in data-driven modeling, and compensate for the lack of labeled data through meaningful priors. In this chapter, we propose a novel approach for audio source separation based on generative priors trained on individual sources. Through the use of projected gradient descent optimization, our approach simultaneously searches in the source-specific latent spaces to effectively recover the constituent sources. Though the generative priors can be defined in the time domain directly we find that using spectral domain loss functions for our optimization leads to good-quality source estimates. We systematically report our observations on a wide variety of audio benchmarks and demonstrate improvements over state-of-the-art unsupervised baselines.

9.1 Problem Setup

Audio source separation, the process of recovering constituent source signals from a given audio mixture, is a key component in downstream applications such as audio enhancement and music information retrieval (Spanias, Painter, and Atti 2006; Spanias, July 2015). Typically formulated as an inverse optimization problem, source separation has been traditionally solved using a broad class of matrix factorization methods (Makino et al. 2004; Karhunen, Wang, and Vigario 1995; Thiagarajan, Ramamurthy, and Spanias 2013), e.g., Independent Component Analysis (ICA) and Principal Component Analysis (PCA). While these methods are known to be effective in over-determined scenarios, i.e. the number of mixture observations is greater than the number of sources, they are severely challenged in under-determined settings (Wang, Reiss, and Cavallaro 2016). Consequently, in the recent years, supervised deep learning based solutions have become popular for under-determined source separation (Stoller, Ewert, and Dixon 2018; Luo and Mesgarani 2019; Lluis, Pons, and Serra 2018; Takahashi, Goswami, and Mitsufuji 2018; Grais, Ward, and Plumbley 2018; Défossez et al. 2019). These approaches can be broadly classified into time domain and spectral domain methods, and often produce state-of-the-art performance on standard benchmarks. Despite their effectiveness, there is a fundamental drawback with supervised methods. In addition to requiring access to large number of observations, a supervised source separation model is highly specific to the given set of sources and the mixing process, consequently requiring complete re-training when those assumptions change. This motivates a strong need for the next generation of unsupervised separation methods that can leverage the recent advances in data-driven modeling, and compensate for the lack of labeled data through meaningful priors.

Utilizing appropriate priors for the unknown sources has been an effective approach to regularize the ill-conditioned nature of source separation. Examples include non-Gaussianity, statistical independence, and sparsity (Virtanen 2003). With the emergence of deep learning methods, it has been shown that choice of the network architecture implicitly induces a structural prior for solving inverse problems (Ulyanov, Vedaldi, and Lempitsky 2018). Based on this finding, Tian *et al.* recently introduced a *deep audio prior* (DAP) (Tian, Xu, and Li 2019) that directly utilizes the structure of a randomly initialized neural network to learn time-frequency masks that isolate the individual components in the mixture audio without any pre-training. Interestingly, DAP was shown to outperform several classical priors.

Here, we consider an alternative approach for under-determined source separation based on *data priors* defined via deep generative models, and in particular using generative adversarial networks (GANs) (Goodfellow et al. 2014). We hypothesize that such a data prior will produce higher quality source estimates by enforcing the estimated solutions to belong to the data manifold. While GAN priors have been successfully utilized in inverse imaging problems (Bora et al. 2017; Zhu et al. 2017; Shah and Hegde 2018; Anirudh et al. 2020) such as denoising, deblurring, compressed recovery etc., their use in source separation has not been studied yet – particularly in the context of audio. In this work, we propose to utilize GAN priors to solve the problem of under-determined source separation. Existing solutions with data priors utilize a single GAN model to perform the inversion process (Anirudh et al. 2020). However, by design, source separation requires the simultaneous estimation of multiple disparate source signals. While one can potentially build a generative model that can jointly characterize all sources, it will require significantly large amounts of data. Hence, we advocate the use of source-specific generative models and generalizing

the PGD optimization with multiple GAN priors. In addition to reducing the data needs, this approach provides the crucial flexibility of handling new sources, without the need for retraining the generative models for all sources. From our study, we find that utilizing multiple GAN priors $\{\mathcal{G}_i | i = 1 \dots K\}$ to be highly effective for under-determined source separation. In particular, we choose a popular waveform synthesis model WaveGAN (Donahue, McAuley, and Puckette 2019) as our GAN prior $\mathcal{G}_i$ as we found the generated samples to be of high perceptual quality. While we utilize time domain GAN prior models, we find that spectral domain loss functions are critical in source estimation using PGD. Using standard benchmark datasets (spoken digit audio (SC09), drums and piano), we evaluate our approach under the assumption that mixing process is known. From our rigorous empirical study, we find that our *data prior* is consistently superior to other commonly adopted priors, including the recent deep audio prior (Tian, Xu, and Li 2019).

9.2 Approach

Audio source separation involves the process of recovering constituent sources $\{\mathbf{s}_i \in \mathbb{R}^d | i = 1 \cdots K\}$ from a given audio mixture $\mathbf{m} \in \mathbb{R}^d$, where K is the total number of sources and d is the number of time steps. In this work, without loss of generality, we assume the source and mixtures to be mono-channel and the mixing process to be a sum of sources i.e., $\mathbf{m} = \sum_{i=1}^{K} \mathbf{s}_i$. Figure 39 provides an overview of our approach for unsupervised source separation. Here, we sample the source audio from the respective priors and perform additive mixing to reconstruct the mixture *i.e*, $\hat{\mathbf{m}} = \sum_{i=1}^{K} \mathcal{G}_i(\mathbf{z}_i)$. The mixture is then processed to obtain the corresponding spectrogram. In addition, we also compute the source level spectrograms. We perform source separation by

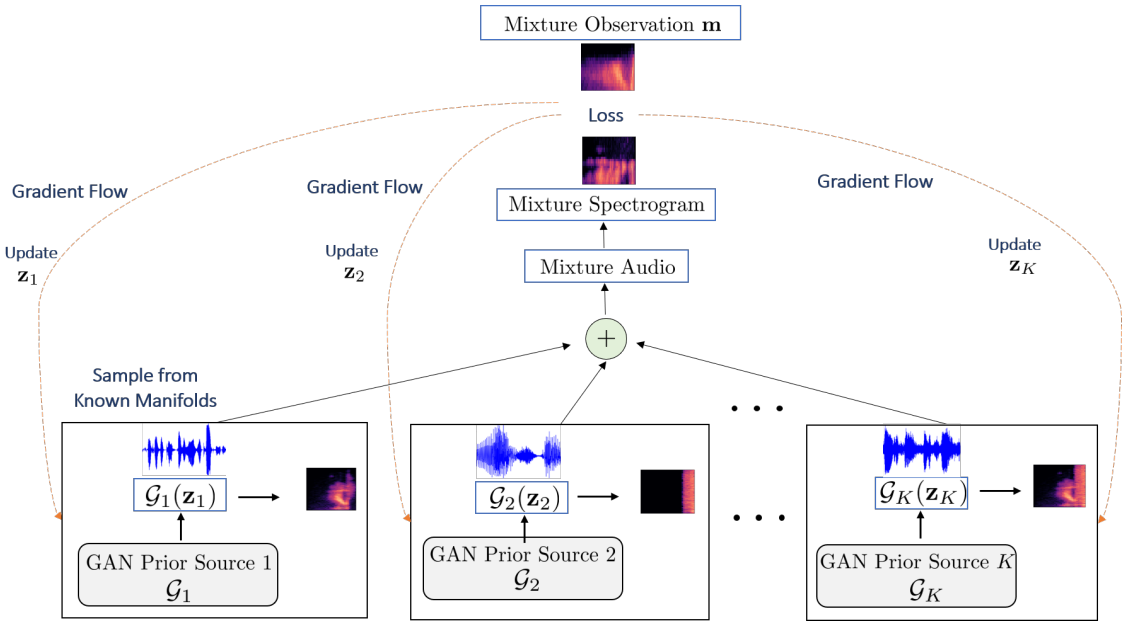


Figure 39. An Overview of the Proposed Unsupervised Source Separation System.

efficiently searching the latent space of the source-specific priors $\mathcal{G}_i$ using *Projected Gradient Descent* optimizing a spectral domain loss function $\mathcal{L}$. More formally, for a single mixture $\mathbf{m}$, our objective function is given by,

$$\{\mathbf{z}_i^*\}_{i=1}^K = arg \min_{\mathbf{z}_1,\mathbf{z}_2 \ldots \mathbf{z}_K} \mathcal{L}(\hat{\mathbf{m}}, \mathbf{m}) + \mathcal{R}(\{\mathcal{G}_i(\mathbf{z}_i)\}), \tag{9.1}$$

where the first term measures the discrepancy between the true and estimated mixtures and the second term is an optional regularizer on the estimated sources. In every PGD iteration, we perform a projection $\mathcal{P}$, where we constrain the $\{\mathbf{z}_i\}_{i=1}^K$ to their respective manifolds. Upon completion of this optimization, the sources can be obtained as $\hat{\mathbf{s}}_i^* = \mathcal{G}_i(\mathbf{z}_i^*), \forall i$. Here, we reformulate the process of source separation by first estimating the source-specific latent features $\mathbf{z}_i^*$ followed by sampling from the respective generators. There are two key ingredients that are critical to the performance of our approach: (i) choice of a good quality *GAN Prior* for every source

Algorithm 1: Approach.

Input: Unlabeled mixture $\mathbf{m}$, No. of sources K,
Pre-trained *GAN Priors* $\{\mathcal{G}_i\}_{i=1\ldots K}$
Output: Estimated sources $\{\hat{\mathbf{s}}_i^*\}_{i=1\ldots K}$
Initialization: $\{\hat{\mathbf{z}}_i\}_{i=1\ldots K} = \mathbf{0} \in \mathbb{R}^{d_z}$
for $t \leftarrow 1$ **to** T **do**
  $\hat{\mathbf{m}} = \sum_{i=1}^{K} \mathcal{G}_i(\hat{\mathbf{z}}_i)$
  Compute source level and mixture spectrograms
  Compute loss $\mathcal{L}$ using 9.6
  $\hat{\mathbf{z}}_i \leftarrow \hat{\mathbf{z}}_i - \eta \nabla_z(\mathcal{L}) \qquad \forall i = 1 \ldots K$
  $\hat{\mathbf{z}}_i \leftarrow \mathcal{P}(\hat{\mathbf{z}}_i)$ $\mathcal{P}$ projects $\{\mathbf{z}_i\}_{i=1\ldots K}$ onto the manifold, i.e., clipped to $[-1, 1]$
end
return $\{\hat{\mathbf{s}}_i^*\} = \mathcal{G}_i(\mathbf{z}_i^*), \forall i$

and (ii) carefully chosen loss functions to drive the PGD optimization. We now elaborate our methodology in the rest of this section.

9.2.1 WaveGAN for Data Prior Construction

WaveGAN (Donahue, McAuley, and Puckette 2019) is a popular generative model capable of synthesizing raw waveform audio. It has exhibited success in producing audio from different domains such as speech and musical instruments. Both the generator and discriminator of the WaveGAN model are similar in construction to DCGAN (Radford, Metz, and Chintala 2015) with certain architectural changes to support audio generation. The generator $\mathcal{G}$ transforms the latent features $\mathbf{z} \in \mathbb{R}^{d_z}$ where $d_z = 100$ from a uniform distribution in $[-1, 1]$, to produce waveform audio $\mathcal{G}(\mathbf{z})$ of dimension $d = 16384$ which is approximately of 1s duration at a sampling rate of 16kHz. The discriminator $\mathcal{D}$ regularized using phase shuffle learns to distinguish between the real and synthesized samples. The WaveGAN is trained to optimize

Table 11. Performance Metrics Averaged Across 1000 Cases for the Digit-piano (k = 2) Experiment (While Higher Spectral SNR and SIR Are Better, Lower RMS Env.Distance Is Better).

Method	**Spectral SNR (dB)**		**RMS Env. Distance**		**SIR (dB)**	
	Digit	**Piano**	**Digit**	**Piano**	**Digit**	**Piano**
FastICA	-2.13	-13.45	0.22	0.61	-4.12	-0.66
PCA	-2.04	-12.01	0.22	0.54	-4.13	-1.44
Kernel PCA	-2.04	-3.30	0.22	0.26	-4.13	-1.61
NMF	-2.21	-5.80	0.23	0.26	-4.09	2.53
DAP	-1.77	**2.72**	0.22	0.22	2.20	-3.10
Proposed	**1.06**	**2.73**	**0.17**	**0.21**	**3.91**	**8.57**

Table 12. Performance Metrics Averaged Across 1000 Cases for the Drums-Piano (K = 2) Experiment.

Method	**Spectral SNR (dB)**		**RMS Env. Distance**		**SIR (dB)**	
	Drums	**Piano**	**Drums**	**Piano**	**Drums**	**Piano**
FastICA	-5.25	-13.52	0.24	0.61	-6.51	-1.45
PCA	-5.19	-12.33	0.24	0.56	-6.53	-2.69
Kernel PCA	-5.19	-3.36	0.24	0.25	-6.53	-2.02
NMF	-5.39	-5.84	0.24	0.26	-6.59	3.84
DAP	-4.20	2.97	0.22	**0.21**	-21.62	**11.22**
Proposed	**0.84**	**3.06**	**0.10**	**0.21**	**11.70**	9.80

the Wasserstein loss with gradient penalty (WGAN-GP) as prescribed in (Gulrajani et al. 2017).

Given the ability of WaveGAN to synthesize high quality audio, the pre-trained generator of WaveGAN was used to define the *GAN Prior*. In our formulation, instead of using a single *GAN Prior* trained jointly for all sources, we construct K independent source-specific priors.

Table 13. Performance Metrics Averaged Across 1000 Cases for the Digit-Drums ($K = 2$) Experiment.

Method	**Spectral SNR (dB)**		**RMS Env. Distance**		**SIR (dB)**	
	Digit	**Drums**	**Digit**	**Drums**	**Digit**	**Drums**
FastICA	2.91	-21.01	**0.13**	0.82	3.10	0.09
PCA	2.99	-20.00	**0.13**	0.77	3.12	0.02
Kernel PCA	2.99	-10.53	**0.13**	0.35	3.12	0.85
NMF	3.01	-13.75	**0.13**	0.39	3.20	-0.98
DAP	**3.59**	**0.92**	0.14	0.14	4.24	-11.48
Proposed	2.32	0.42	0.15	**0.10**	**25.91**	**23.68**

9.2.2 Losses

In order to obtain high-quality source estimates using GAN priors, we propose a novel yet intuitive combination of spectral-domain losses. Though one can utilize time-domain metrics such as the Mean-Squared Error (MSE) to compare the observed and synthesized mixtures, we find that even small variations in the phases of sources estimated from our priors can lead to higher error values. This in turn can misguide the PGD optimization process and may lead to poor convergence. This corroborates with the findings in (Défossez et al. 2018).

9.2.2.1 Multiresolution Spectral Loss ($\mathcal{L}_{ms}$)

This loss term measures the ℓ_1-norm between log magnitudes of the reconstructed spectrogram and the input spectrogram at L spatial resolutions. This is used to enforce perceptual closeness between the two mixtures at varying spatial resolutions. Denoting $\mathbf{m}$ as the input mixture and $\hat{\mathbf{m}}$ as the estimated mixture, the loss $\mathcal{L}_{ms}$ is

defined as

$$\mathcal{L}_{ms} = \sum_{l=1}^{L} \Big\| \log\big(1 + |STFT^{l}(\mathbf{m})|^2\big) - \log\big(1 + |STFT^{l}(\hat{\mathbf{m}})|^2\big) \Big\|_1, \tag{9.2}$$

where $|STFT^{l}(.)|$ represents the magnitude spectrograms at the l^{th} spatial resolution and $L = 3$. We compute the magnitude spectrogram at different resolutions by performing a simple average pooling operation with bilinear interpolation.

9.2.2.2 Source Dissociation Loss ($\mathcal{L}_{sd}$)

Minimizing this loss, defined as the aggregated gradient similarity between the spectrograms of the estimated sources, enforces them to be systematically different. Similar to (Tian, Xu, and Li 2019; Zhang, Ng, and Chen 2018), we define this as a product of the normalized gradient fields of the log magnitude spectrograms computed at L spatial resolutions. In the case where there are K constituent sources, we compute this between every pair of sources. Formally,

$$\mathcal{L}_{sd} = \sum_{i=1}^{K} \sum_{j=i+1}^{K} \sum_{l=1}^{L} ||\Psi(\log\big(1 + |STFT^{l}(\mathcal{G}_i(\hat{\mathbf{z}}_i))|^2\big), \log\big(1 + |STFT^{l}(\mathcal{G}_j(\hat{\mathbf{z}}_j))|^2\big))||_F, \tag{9.3}$$

where $\Psi(x, y) = tanh(\lambda_1|\nabla x|) \odot tanh(\lambda_2|\nabla y|)$. ($\odot$ represents element-wise multiplication) and $L = 3$. The weights λ_1 and λ_2 are set at $\lambda_1 = \frac{\sqrt{|\nabla y|_F}}{\sqrt{|\nabla x|_F}}$ and $\lambda_2 = \frac{\sqrt{|\nabla x|_F}}{\sqrt{|\nabla y|_F}}$.

9.2.2.3 Mixture Coherence Loss ($\mathcal{L}_{mc}$)

Along with $\mathcal{L}_{ms}$, this loss, defined using gradient similarity between original and reconstructed mixtures, ensures that our PGD optimization produces meaningful reconstructions:

$$\mathcal{L}_{mc} = -\sum_{l=1}^{L} ||\Psi(\log\big(1 + |STFT^l(\mathbf{m})|^2\big), \log\big(1 + |STFT^l(\hat{\mathbf{m}})\big)|^2))||_F \tag{9.4}$$

9.2.2.4 Frequency Consistency Loss ($\mathcal{L}_{fc}$)

This loss helps improve perceptual similarity between the magnitude spectrograms of the input and synthesized mixtures by constraining components within a particular temporal bin of the spectrograms to remain consistent over the entire frequency range, i.e.,

$$\mathcal{L}_{fc} = \sum_{t=1}^{T}\sum_{f=1}^{F} \frac{\log(1 + |STFT(\mathbf{m})[t, f]|)}{\log(1 + |STFT(\hat{\mathbf{m}})[t, f]|)}. \tag{9.5}$$

The overall loss function for our source separation algorithm is thus obtained as:

$$\mathcal{L} = \beta_1\mathcal{L}_{ms} + \beta_2\mathcal{L}_{sd} + \beta_3\mathcal{L}_{mc} + \beta_4\mathcal{L}_{fc} \tag{9.6}$$

Through hyperparameter search we identified that $\beta_1 = 0.8, \beta_2 = 0.3, \beta_3 = 0.1, \beta_4 = 0.4$ to be effective in our experiments. Note, in our computations we obtain the spectrograms by computing the Short Time Fourier Transform (STFT) on the waveform in frames of length 256, hop size of 128 and FFT length of 256. The procedure for our approach is showed in Algorithm 1. Figure 40 illustrates the progressive estimation of the unknown sources using our approach.

Table 14. Performance Metrics Averaged Across 1000 Cases for the Digit-Drums-Piano ($K = 3$) Experiment.

Metric	Source	FastICA	PCA	Kernel PCA	NMF	Proposed
Spectral SNR (dB)	Digit	-2.95	-2.47	-2.47	-2.47	**0.77**
	Drums	-10.8	-19.81	-8.1	-12.84	**0.64**
	Piano	0.27	0.1	-0.94	**4.94**	2.64
RMS Env. Distance	Digit	0.24	0.23	0.23	0.23	**0.17**
	Drums	0.4	0.75	0.28	0.37	**0.1**
	Piano	0.23	0.31	0.25	**0.15**	0.21
SIR (dB)	Digit	-4.73	-5.06	-5.06	-5.01	**3.02**
	Drums	-6.48	-5.51	-1.65	-5.69	**10.21**
	Piano	0.53	2.21	-3.87	2.60	**5.12**

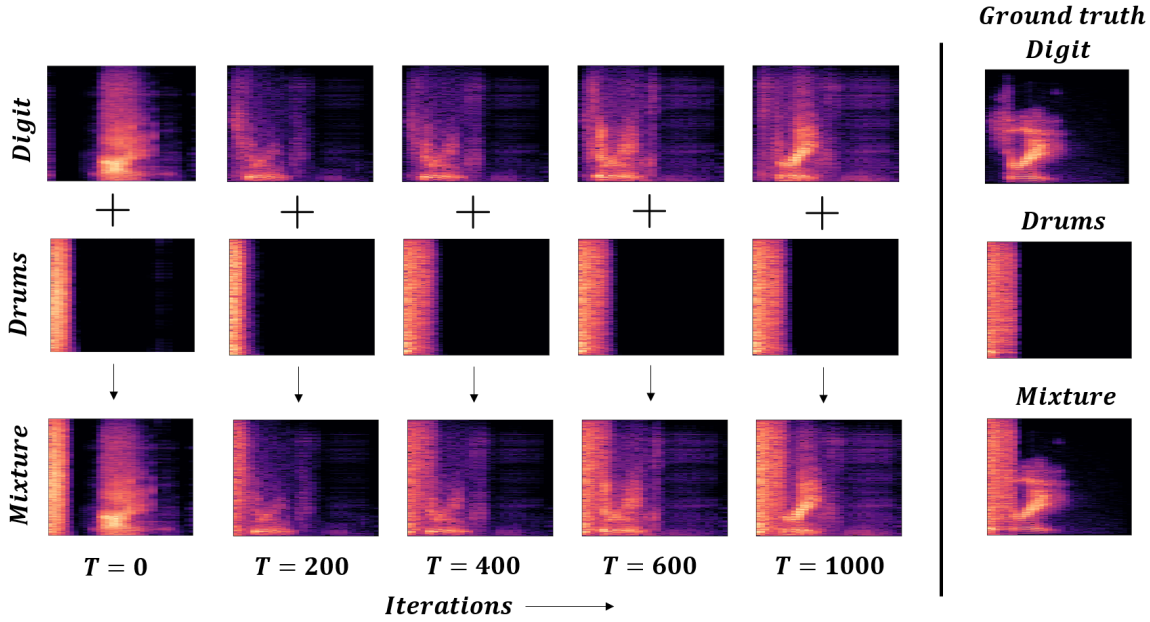


Figure 40. Demonstration of Our Approach Using a Digit-drum Example. Through the Use of Multiple *GAN Priors* $\mathcal{G}_i$, Our Algorithm Efficiently Searches the Source-specific Latent Spaces to Estimate the Underlying Sources.

In this section, we evaluate our approach on two source and three source separation experiments on the publicly available Spoken Digit (SC09), drum sounds and piano datasets. The SC09 dataset is a subset of the Speech Commands dataset (Warden 2018; Donahue, McAuley, and Puckette 2019) containing spoken digits (0-9) each of duration $\sim$ 1s at 16kHz from a variety of speakers recorded under different acoustic conditions. The drum sounds dataset (Donahue, McAuley, and Puckette 2019) contains single drum hit sounds each of duration $\sim$ 1s at 16kHz. The piano dataset (Donahue, McAuley, and Puckette 2019) contains piano music (Bach compositions) each of duration ($>$ 50s) at 48kHz.

WaveGAN Training. Following (Donahue, McAuley, and Puckette 2019), we train WaveGAN models on normalized 1s slices (*i.e* d =16384 samples) of the SC09 (Digit), Drums and Piano train datasets resampled to 16kHz respectively. All the models were trained using batches of size 128. The generator and discriminator were optimized using the WGAN-GP loss with the Adam optimizer and learning rate $1e^{-4}$ for 3000 epochs. The trained generator models were used to construct the GAN priors.

Setup. For the task of two source separation ($K = 2$), we conducted experiments on three possible mixture combinations: (i) Digit-Piano, (ii) Drums-Piano and (iii) Digit-Drums. In order to create the input mixture for every combination, we randomly sampled (with replacement) normalized 1s audio slices from the respective test datasets, and obtained 1000 mixtures through a simple additive mixing process. Similarly, we obtained 1000 mixtures for the case of $K = 3$, i.e., on the combination, Digit-Drums-Piano. In each case, we performed the PGD optimization using Eqn.9.6 for 1000 iterations with the ADAM optimizer and learning rate of $5e^{-2}$ to infer

source specific latent features $\{\mathbf{z}_i^*\}_{i=1\ldots K}$. The estimated sources are then obtained as $\{\mathcal{G}_i(\mathbf{z}_i^*)\}_{i=1\ldots K}$. Though the choice of initialization for $\mathbf{z}_i$ is known to be critical for PGD optimization (Anirudh et al. 2020), we find that setting $\{\mathbf{z}_i\}_{i=1\ldots K} = \mathbf{0} \in \mathbb{R}^{d_z}$ to be effective.

Evaluation Metrics. Following standard practice, we used three different metrics - (i) mean spectral SNR (Spiertz and Gnann 2009; Virtanen 2007), a measure of the quality of the spectrogram reconstruction; (ii) mean RMS envelope distance (Morgado et al. 2018) between the estimated and true sources; and (iii) mean signal-interference ratio (SIR) (Stöter, Liutkus, and Ito 2018) to quantify the interference caused by one estimated source on another.

Results. Tables 11, 12, 13 and 14 provide a comprehensive comparison of our approach against the standard baselines (FastICA, PCA, KernelPCA, NMF) (Pedregosa et al. 2011) as well as with the state-of-the-art unsupervised Deep-Audio-Prior (Tian, Xu, and Li 2019). It can be observed that our approach significantly outperforms all the baselines in most cases, except for the Digits-Drums experiment where our method is in par with DAP. These results indicate the effectiveness of our unsupervised approach on complex source separation tasks. We find that the spectral SNR metric, which is relatively less sensitive to phase differences (Défossez et al. 2018; Spiertz and Gnann 2009), is consistently high with our approach, indicating high perceptual similarities between estimated and the ground truth audio. We also find lower envelope distance estimates, further emphasizing the perceptual quality of our estimated sources. Finally, we attribute the significant improvements in the SIR metric to the *source dissociation loss* ($\mathcal{L}_{sd}$), which enforces the estimated sources from the priors to be systematically different.

9.4 Summary

In summary, we found that source-specific *GAN Priors* are effective in recovering the constituents of an unlabeled mixture, often significantly outperforming unsupervised state-of-the-art benchmarks. Additionally, we found that such generative priors can be further improved with PGD-style optimization using carefully designed spectral domain loss functions. Our approach is highly flexible because it is entirely an inference-time technique, and as a result can efficiently deal with varying number of known sources in a given mixture. This was in contrast with standard supervised approaches which require re-training or extensive fine-tuning. Future extensions to our work include performing source separation when the mixing process is unknown, and dealing with mixtures that contain novel sources.

Chapter 10

DESIGN OF DEEP MODEL PRIORS FOR UNSUPERVISED AUDIO RESTORATION

In this chapter, the problem of designing a deep model prior architecture for solving unsupervised audio restoration is discussed. We identify the gaps in the existing strategies of prior designs and propose a novel yet efficient prior design that produces high fidelity audio restoration. The critical impact of introducing such an architecture to handle the statistical variations in the nature of audio signals are also elucidated.

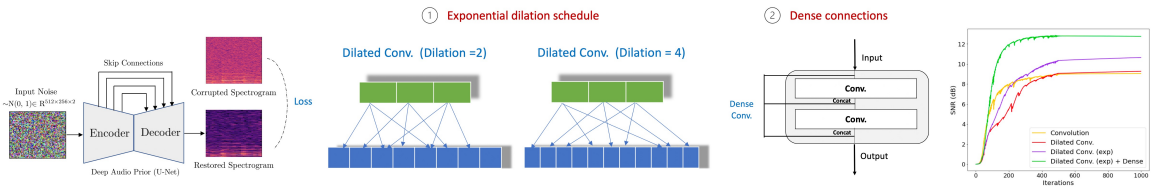
10.1 Problem Setup

Deep convolutional neural networks (CNNs) have proven to be effective for recovering signals from noisy observations (Anirudh et al. 2020). Consequently, state-of-the-art solutions for challenging problems such as audio enhancement (Pascual, Bonafonte, and Serra 2017), audio inpainting (Y.-L. Chang et al. 2019) and source separation (Luo and Mesgarani 2019) are based on convolutional architectures (Giri, Isik, and Krishnaswamy 2019; Défossez et al. 2019). While majority of this success has been with supervisory data, recent focus has shifted to unsupervised approaches that do not require expensive data collection and curation. Given the ill-posed nature of audio restoration, choice of suitable audio priors is critical to the success of unsupervised learning approaches.

The seminal work of Ulyanov *et al.* (Ulyanov, Vedaldi, and Lempitsky 2018) introduced the notion of *deep image priors* and showed that convolutional neural

network architectures can provide powerful signal priors for solving image restoration problems. In contrast to unsupervised approaches that use priors based on pre-trained generative models (e.g, Generative Adversarial Networks (GANs)) (Shah and Hegde 2018; Vivek Narayanaswamy et al. 2020) to solve inverse problems, these model or structural priors do not require any training data and the optimization can be carried out using a single observation. The flexibility and the effectiveness of this approach has motivated the design of suitable priors for audio restoration tasks. A number of recent studies (Z. Zhang et al. 2019; Tian, Xu, and Li 2019; Michelashvili and Wolf 2019; Y.-L. Chang et al. 2019) have showed that different variants of convolutional architectures are highly effective choices. For example, Michelashvili et.al (Michelashvili and Wolf 2019) used the Wave-U-Net (Stoller, Ewert, and Dixon 2018) architecture to denoise audio signals. Interestingly, convolutional network constructions that operate in the spectral domain have been found to be consistently superior. For example, Tian *et al.* (Tian, Xu, and Li 2019) proposed Deep Audio Priors, that utilize separate randomly initialized U-Net models (Ronneberger, Fischer, and Brox 2015) to obtain time-frequency masks and audio source estimates respectively for source separation without any pre-training.

Deep audio priors can be characterized using a number of factors including recovery performance across different inversion tasks, ease of training, and computational efficiency. For example, replacing standard convolutions with dilated convolutions (Oord et al. 2016; Yu and Koltun 2015) is known to improve recovery performance without any impact on the computational efficiency. More recently, Zhang *et al.* (Z. Zhang et al. 2019) explored the use of harmonic convolutions that carefully engineer the convolutional kernels to better capture multi-scale harmonic structure in audio. Takeuchi *et al.* (Takeuchi et al. 2020) subsequently improved the computational efficiency of



(a) Unsupervised restoration using deep audio priors (b) Proposed modifications to U-Net based audio priors (c) Performance comparison

Figure 41. We Propose a New Deep Audio Prior Construction That Is Well Suited for Challenging Unsupervised Restoration Tasks. Through the Use of Dilated Convolutions with a Carefully Engineered Dilation Schedule and Dense Connections in a Standard U-net, We Achieve Significant Performance Gains over State-of-the-art Approaches. The Example in (c) Corresponds to an Audio Denoising Experiment.

harmonic convolutions through harmonic lowering. However, audio priors based on harmonic convolutions require significantly larger number of training iterations compared to standard convolutions. While this challenge was addressed in (Z. Zhang et al. 2019) through the use of multiple anchors for harmonic convolutions, the resulting audio prior can be of significantly higher complexity.

In this work, we revisit the design of audio priors (spectral domain) and propose a new U-Net based prior, that does not impact either the network complexity or the convergence behavior, but consistently leads to high-fidelity restoration. We find that unsupervised audio restoration can be improved by adopting dilated convolutions with an exponentially increasing dilation schedule and by introducing dense connections. We perform extensive empirical studies on audio denoising, inpainting and source separation and demonstrate that our audio prior better extracts multi-scale features from time-frequency representations of audio signals, significantly outperforms widely adopted deep audio priors, and is computationally efficient when compared to harmonic convolutions (Z. Zhang et al. 2019; Takeuchi et al. 2020).

10.2 Unsupervised Audio Restoration

Audio restoration refers to the process of recovering an audio signal $\mathbf{x} \in \mathbb{R}^{m \times c}$ from a corrupted observation $\hat{\mathbf{x}} \in \mathbb{R}^{n \times c}$ (Spanias, Painter, and Atti 2006). Here m and n denote the length of the observations while c denotes the number of channels. Without loss of generality, in this work, we assume $m = n$ and the signals to be mono-channel i.e., $c = 1$. In this work, we consider three popular audio restoration tasks, namely audio denoising, inpainting and source separation.

Audio denoising refers to the task of removing noise from a corrupted audio while preserving the underlying characteristics. On the other hand, audio inpainting attempts to recover the original signal from observations that are spatio-temporally masked, and is typically utilized for audio editing and packet loss recovery in receiver systems. Finally, source separation refers to the process of recovering the constituent audio sources present in a given mixture observation, wherein the mixing process may not be known in advance.

In practice, since the corruption process (e.g., type of noise or noise level) is unknown *a* priori, audio restoration is a severely ill-posed inverse problem and often requires meaningful signal priors (Ulyanov, Vedaldi, and Lempitsky 2018; Z. Zhang et al. 2019). In this context, deep audio priors have become highly prevalent, particularly for unsupervised restoration. Formally, given a corrupted observation $\hat{\mathbf{x}}$ and an untrained convolutional neural network f_Θ with parameters Θ, the structure of the neural network can innately regularize the inverse optimization. The intuition behind such structural priors is that if the network is capable of modeling the signal priors induced by its structure, the network would fit the signal easily than that noise. In a

deep audio prior-based restoration, the clean signal can be directly obtained as $f_\Theta(\mathbf{z})$, where $\mathbf{z}$ is a random noise (latent) vector drawn from a known distribution.

In this work, we study the design of effective deep audio priors for practical restoration tasks. Though existing efforts in the literature have explored the use of dilated and harmonic convolutions in U-Net based priors, large performance gains and desirable training behavior were enabled only by increasing the complexity of the prior, e.g., multiple anchors for harmonic convolutions (Z. Zhang et al. 2019). In contrast, we propose key modifications to U-Net based audio priors that do not significantly increase the network complexity, but can produce large performance gains in restoration tasks. More specifically, we propose the use of exponentially increasing receptive fields via dilated convolutions (Oord et al. 2016) by adopting a pre-specified dilation schedule which dispenses the need for explicit resampling techniques for better feature extraction. Furthermore, we introduce dense connections between within each layer, as well as between *upstream* and *downstream* paths of the U-Net to promote better feature reuse and improved gradient flows. Together, deep audio priors with these two modifications consistently outperform other widely adopted prior choices.

10.3 Approach

Figure 41 provides an overview of our approach. We propose a new U-Net based deep audio prior construction that we empirically find to be superior to existing convolutional architectures for unsupervised restoration. In this section, we describe the key steps of our approach: (i) designing an U-Net architecture; (ii) using dilated convolutions with a specific dilation schedule; and (iii) adding dense connections for improved gradient flow.

10.3.1 U-Net Architecture Design

We adopt the U-Net architecture as a structural prior to effectively regularize the ill-posed tasks of audio restoration. The architecture is comprised of two convolutional blocks in the *downstream* path where each block in turn contains two 2D convolution layers with filter sizes $\{35, 70\}$ and $\{70, 140\}$ respectively. Correspondingly, the *upstream* path is comprised of two convolutional blocks, wherein each block contains a bi-linear upsampling step followed by two convolution layers with filter sizes $\{140, 70\}$ and $\{70, 35\}$ respectively. The bottleneck block between *downstream* and *upstream* paths consists of two more convolutional layers with 70 filters each. The final output is obtained using another convolutional layer with the desired number of channels. In addition, skip connections are included between the convolutional blocks in the downstream and upstream paths, which combine the coarse and fine grained features from the respective paths to improve signal reconstruction.

10.3.2 Dilated Convolutions with an Exponential Schedule

The success of the audio prior relies heavily on the quality of the features extracted at different scales for signal reconstruction. Recovering audio signals can be challenging due to the inherent periodicities and complex spatio-temporal statistics, and this naturally motivates feature extraction strategies that can leverage information over wider receptive fields at increasing depths. To this end, we introduce dilations in all convolutional layers of the U-Net, wherein the dilation rates are exponentially increased for each subsequent convolution layer (in factors of 2). Specifically, starting with a dilation factor of 2 for the first convolution layer in the first block, the dilation

rate grows upto 32 in the bottleneck block. The *upstream* is correspondingly designed to mirror the *downstream* architecture. The inherent downsampling operation in the *downstream* path (max-pooling) combined with the exponential dilation schedule effectively enables feature extraction across significantly large receptive fields (e.g., periodicities).

10.3.3 Adding Dense Connections

In addition to enabling multi-scale feature extraction via an exponential dilation schedule, we aim to enhance the U-Net architecture further by adding dense connections in order to encourage feature reuse and improve gradient flow even at increasing layer depths (see Fig.41(b)). More specifically, we include dense connections between convolutional layers within each convolutional block, i.e., the feature maps produced by each layer are concatenated to the subsequent layers in the block. In order to prevent the accumulation of a large number of feature maps at increasing depths, following Thiagarajan *et al.* (Thiagarajan, Rajan, et al. 2020), we include a transition block (implemented using a single standard $2-$D convolutional layer) at the end of every dense block, which reduces the dimensionality of the resulting feature maps.

Comparison to Harmonic Convolutions. Recent efforts (Z. Zhang et al. 2019) have recommended harmonic convolutions as an effective choice over standard convolutions for designing audio priors. However, as illustrated in Figure 42 for an audio denoising example, our audio prior construction requires significantly lower number of iterations ($0.1\times$) to converge when compared to harmonic convolutions, while also providing non-trivial gains in the restoration performance.

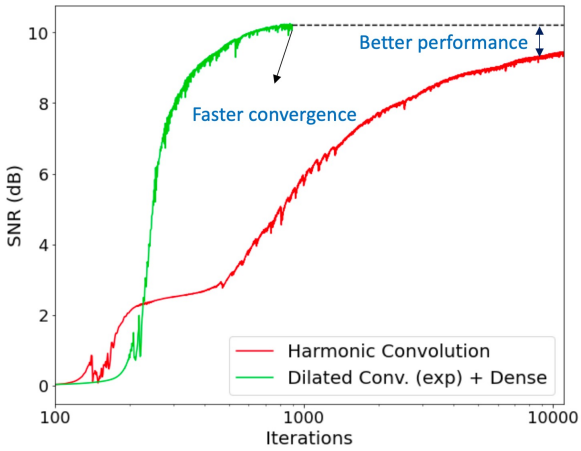


Figure 42. Comparing the Convergence of Our Audio Prior to U-nets Based on Harmonic Convolutions. Our Prior Achieves Both Significantly Faster Convergence and Marginal Performance Gains over the Latter, While Convincingly Outperforming Other Widely Adopted Deep Audio Prior Constructions.

Table 15. Audio Denoising Performance of Deep Audio Priors under the Presence of Gaussian Noise.

DAP Design	LJ-Speech		Digits		Piano
	PESQ	PSNR	PESQ	PSNR	PSNR
Convolution	1.73 ± 0.17	6.85 ± 1.47	1.99 ± 0.47	9.54 ± 2.96	9.17 ± 1.34
Dilated Conv.	1.76 ± 0.21	7.18 ± 1.55	2.08 ± 0.47	10.85 ± 2.72	9.11 ± 1.49
Dilated Conv. (exp)	2.00 ± 0.19	7.68 ± 1.76	2.20 ± 0.45	11.52 ± 2.89	10.68 ± 1.47
Dilated Conv. (exp) + Dense	**2.07 ± 0.16**	**8.15 ± 1.94**	**2.23 ± 0.47**	**11.55 ± 2.91**	**12.50 ± 1.11**

10.4 Experiments

In this section, we present empirical studies to evaluate our approach on three ill-posed audio restoration tasks, namely denoising, in-painting and source separation. We will begin by discussing the datasets used in our study.

Table 16. Audio Denoising Performance of Deep Audio Priors under the Presence of Environmental Noise.

DAP Design	LJ-Speech		Digits		Piano
	PESQ	PSNR	PESQ	PSNR	PSNR
Convolution	1.91 ± 0.26	4.36 ± 1.29	2.23 ± 0.58	6.39 ± 1.81	5.73 ± 1.06
Dilated Conv. (exp)	2.04 ± 0.24	5.03 ± 1.27	2.31 ± 0.69	6.95 ± 1.92	6.40 ± 1.12
Dilated Conv. (exp) + Dense	**2.31 ± 0.22**	**5.58 ± 1.34**	**2.46 ± 0.58**	**7.19 ± 1.82**	**7.30 ± 1.12**

Table 17. Audio Inpainting Performance of Deep Audio Priors under Random Spatio-temporal Masking.

DAP Design	LJ-Speech		Digits	
	Spec. SNR	Env. Dist.	Spec. SNR	Env. Dist.
Convolution	7.27 ± 1.43	0.08 ± 0.02	7.34 ± 1.91	0.12 ± 0.03
Dilated Conv.	7.18 ± 1.06	0.08 ± 0.02	7.78 ± 1.88	0.11 ± 0.03
Dilated Conv. (exp)	7.95 ± 1.16	0.07 ± 0.02	9.02 ± 1.96	0.10 ± 0.03
Dilated Conv. (exp) + Dense	**10.01 ± 1.78**	**0.06 ± 0.02**	**10.80 ± 2.52**	**0.09 ± 0.03**

Table 18. Unsupervised Source Separation Performance of Deep Audio Priors.

DAP Design	SDR (dB)		SIR (dB)		Spec. SNR (dB)		Env. Dist	
	Piano	Drums	Piano	Drums	Piano	Drums	Piano	Drums
Convolution	2.56 ± 1.71	-0.28 ± 1.45	13.17 ± 6.62	-6.41 ±8.18	2.75 ± 1.59	0.01 ± 0.97	0.28 ± 0.09	0.18 ± 0.07
Dilated Conv.	2.54 ± 1.63	0.02 ± 1.47	13.09 ± 7.47	-3.41 ± 9.09	2.75 ± 1.39	0.04 ± 0.74	0.26 ± 0.08	0.15 ± 0.05
Dilated Conv. (exp)	3.07 ± 1.38	0.15 ± 1.87	11.84 ± 8.64	1.12 ± 5.25	3.26 ± 1.64	0.17 ± 1.65	0.25 ± 0.06	**0.14 ± 0.05**
Dilated Conv. (exp) + Dense	**4.84 ± 2.61**	**0.61 ± 3.09**	**12.57 ± 7.62**	**1.93 ± 5.64**	**5.43 ± 2.04**	**0.54 ± 1.77**	**0.21 ± 0.08**	**0.14 ± 0.04**

10.4.1 Datasets.

We used the following datasets for our study: LJSpeech, SC09 Spoken Digit (SC09), drum and piano sounds. LJSpeech (Ito and Johnson 2017) is an open source dataset

containing audio clips of duration $\sim 8s$ at 22kHz of a speaker reading sentences. The SC09 dataset (Warden 2018; Donahue, McAuley, and Puckette 2019) is comprised of spoken digits (0-9) with duration $\sim 1s$ at 16kHz. The drum sounds dataset (Donahue, McAuley, and Puckette 2019) contains single drum hit audio of duration $\sim 1s$ at 16kHz, while the piano dataset (Donahue, McAuley, and Puckette 2019) contains clips of duration $> 50s$ at 48kHz.

Pre-processing. In all our experiments, we resample the audio samples to 16kHz and use clips of duration $2s$ for LJSpeech and $1s$ for other datasets. We carry then compute the spectrograms for the audio clips, using window length 1022 and hop length 64. Following Zhang (Z. Zhang et al. 2019) *et al.*, we utilize both the real and imaginary parts of the spectrogram as a $2-$channel input.

10.4.2 Baselines

. We compare the performance of our audio prior to the widely adopted U-Net priors based on regular convolutions and dilated convolutions (constant dilation factor). For ablation, we also considered a variation where we used the exponential dilation schedule without dense connections. Though harmonic convolution (Z. Zhang et al. 2019) is another choice for implementing the audio prior, due to its significantly slower convergence (see Figure 42), we did not include it as a baseline approach. However, from our experiments, we found that our approach consistently outperformed U-Nets with harmonic convolutions.

10.5 Performance Evaluation on Audio Restoration Tasks

10.5.1 Audio Denoising

In this task, using single corrupted observation $\hat{\mathbf{x}}$, we use deep audio priors to recover the underlying clean signal $\mathbf{x}$:

$$\min_{\Theta} \mathcal{L}(f_{\Theta}(\mathbf{z}), \hat{\mathbf{x}}),$$

where $f_{\Theta}(\mathbf{z})$ is the restored output from the audio prior parameterized by Θ, and $\mathcal{L}$ is implemented as the ℓ_2 loss. We evaluate our audio prior under two different noise scenarios (i) *Gaussian Noise:* We add Gaussian noise with standard deviation 0.1 to clean audio; (ii) *Environmental Noise:* We used Living Room and Traffic Noise samples from the DEMAND database (Thiemann, Ito, and Vincent 2013) and synthesize observations by adding them with the clean audio at SNRs chosen randomly between 5 and 9dB. We performed the optimization on each observation for 2000 iterations using the ADAM optimizer and learning rate 0.001.

Metrics. Follow standard practice, we used the PESQ (Perceptual Evaluation of Speech Quality) and the PSNR (Peak-Signal to Noise Ratio) metrics.

Findings. Tables 15 and 16 show the performance of our approach against the baseline audio prior constructions on both noise settings. We report the performance metrics obtained on 50 random samples from each of the datasets. We find that our approach provides a significantly superior performance over standard convolutions even under challenging environmental noise conditions. Note that, while dilated convolutions with a constant dilation factor are better than regular convolutions, the exponential dilation schedule improves by a bigger margin.

10.5.2 Audio In-painting

In this task, we use deep audio priors to fill masked regions in the observation $\hat{\mathbf{x}}$ that is spatio-temporally masked with a known mask $\mathbf{m}$:

$$\min_{\Theta} \mathcal{L}\Big((\mathbf{m} \odot f_{\Theta}(\mathbf{z})), \mathbf{m} \odot \hat{\mathbf{x}}\Big)$$

Similar to the denoising experiment, we used the ℓ_2 loss for $\mathcal{L}$ and performed the optimization for 2000 iterations. We construct masked observations by randomly introducing zero masks of duration varying between $0.1s$ to $0.25s$, such that all frequency components within the mask are zeroed out.

Metrics. For evaluation, we used the Spectral-SNR (Spiertz and Gnann 2009; Virtanen 2003), a measure of the quality of spectrogram reconstruction, and the RMS Envelope distance (Morgado et al. 2018).

Findings. Table 17 compares our approach against the baselines, using results from 50 examples in each dataset. It can be observed that, our approach consistently outperforms existing methods (2.5dB improvement in SNR on average) and introduces statistically meaningful patterns into the masked regions. This clearly demonstrates the efficacy of our audio prior.

10.5.3 Source Separation

We address the task of two source separation by adopting a formulation similar to (Gandelsman, Shocher, and Irani 2019) - We use two audio priors aim to reconstruct the constituent sources $\{\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2\}$ and another prior to synthesize a mask $\mathbf{m}$, that can be used to mix the constituent sources to create the mixture audio, $\mathbf{m}\hat{\mathbf{s}}_1 + (1 - \mathbf{m})\hat{\mathbf{s}}_2$. Similar to other previous restoration tasks, we use only a single observation

(underdetermined setting) to estimate the sources. We synthesized 50 mixtures by randomly sampling and combining audio clips from the drums and the piano datasets and adopted loss functions from (Tian, Xu, and Li 2019).

Metrics. We used the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) (Stöter, Liutkus, and Ito 2018), Spectral SNR and the RMS envelope distance metrics for evaluation.

Findings. Table 18 illustrates the performance of the DAP design choices. We find that, without increasing the complexity of the audio prior construction, the performance of U-Net architectures can be significantly improved through the proposed strategies. From our results, the effectiveness of our audio prior design even with challenging inverse problems is clearly evident.

10.6 Summary

In this chapter, we developed a new deep audio prior construction that employs a carefully engineered convolutional architecture to produce significant performance gains in unsupervised audio restoration problems. In particular, we found that audio priors can be vastly improved through dilated convolutions with an exponential dilation schedule and dense connections. While the former strategy effectively increased the receptive fields for feature extraction, the latter supported a more principled learning of multi-scale features. As demonstrated by our experiments a suite of ill-posed audio restoration problems, our approach provided meaningful signal priors to regularize this optimization process.

Chapter 11

CONCLUSIONS AND FUTURE WORK

11.1 Conclusions

The importance of accurate model design and characterization for improving reliability and fidelity across different tasks and when deployed under open-world conditions was described in this dissertation. In particular, approaches to building reliable failure detectors for deep deterministic models, designing well-calibrated OOD detectors that can identify a wide range of semantic and modality shifts, rigorously characterizing model confidence to support actionability by utilizing UQ in deep models, and designing next-gen data and model priors for solving ill-posed unsupervised audio restoration were presented. For each problem, a detailed description including the motivation, modeling and optimization strategies, detailed experiment setups, relevant applications, and inferences were provided.

The key findings of this dissertation can be summarized as follows:

- The criticality of the choice of metric adopted to characterize model confidence for the safe adoption of models under real-world distribution shifts was noted. By identifying loss as a direct measure of generalization error, Direct Error Predictors (DEPs) consistent with the underlying model were carefully constructed. A crucial finding was made that jointly training the classifier together with the DEPs with tailored loss functions is critical for regularization and the production of meaningful loss estimates. This strategy was demonstrated to improve the generalization and outlier rejection capabilities of the underlying classifier. It

was also identified that the DEP can effectively detect distribution shifts, offer pixel-level sensitivities, and effectively identify prototypical examples in datasets.

- It was identified that the DEP can effectively detect data distribution shifts without the requirement of any adversarial training strategies, which naturally allowed the DEP to estimate the features of data that are most relevant to explain a prediction via input masking. Under the lens of Granger causality, the DEP was leveraged as an effective tool for feature importance estimation. It was observed that the DEP was resilient to distribution shifts and significantly outperformed other baselines by large margins over a variety of data modalities.
- It was observed that the DEP can effectively be operated as a hypothesis tester of whether any input sample belongs to the inferred training manifold or not. Such a principle was found to be critical for guiding the ill-posed inversion of counterfactual image synthesis. While the use of strong model priors can produce realistic image synthesis, the use of the DEP was critical to project the counterfactuals onto the actual training manifold and introduce discernible changes in the relevant pixels of the images. Progressive optimization was introduced as a strategy for synthesizing CFs, and a novel classifier discrepancy metric was introduced for evaluation. A systematic study was performed over different choices of priors, kernel density based hypothesis testers, and datasets to demonstrate the efficacy of the approach.
- The importance of calibrating OOD detectors in order to improve model generalization and anomaly detection for medical imaging was demonstrated. While a reliable failure indicator such as the DEP can produce well-calibrated uncertainties, it may not be effective in accurately scoring inlier and outlier data. To calibrate OOD detectors, a dual objective was formulated of not compromising

on inlier accuracy while simultaneously rejecting OOD examples based on a scoring metric such as energy. To satisfy this dual objective, regimes of inlier and outlier data were specified through suitable augmentation strategies. It was demonstrated that the choice and space of augmentation plays a critical role in improving OOD detection. In particular, it was identified that specifying inliers in the latent space of the detector and outliers in the pixel space as synthetic corruptions of the training data can lead to improved OOD detectors that can detect a variety of semantic and modality shifts.

- The estimation of aleatoric and epistemic uncertainties that arise in the DNN training pipeline allows for the rigorous characterization of confidence and enables actionability. Under the UQ framework, it was identified that existing approaches for estimating epistemic uncertainty were computationally complex or intractable. A principled finding was made that injecting trivial biases (anchors) onto a dataset and training such an ensemble of models produces inconsistencies in predicting the outcome, which can be considered as the epistemic uncertainty providing signals about data regimes not represented or sampled from. By rolling such an anchor ensemble into a single model, a new training strategy, Δ-UQ, a scalable predictive uncertainty estimator, was designed. It was found that the uncertainties produced by Δ-UQ are significantly superior and outperform existing methods on tasks such as sequential optimization and outlier rejection.
- In addition to using Δ-UQ as a predictive uncertainty estimator, it was found that this principle can also be used for estimating representation uncertainty. It was identified that Δ-UQ implicitly induces a metric characterizing different distributions, a principle that was leveraged to predict the generalization gap in deep models. When compared to state-of-the-art baselines that use summary

metrics, Δ-UQ was found to produce accurate estimates of generalization gaps even under complex distribution shifts.

- The idea of extending a PGD-style, single generative prior based inversion strategy to multiple generative prior based inversion was demonstrated. In particular, it was shown that source-specific generative priors are effective in recovering the constituents of an unlabeled mixture for the ill-posed task of source separation. It was also found that solving source separation with such priors requires the careful design of spectral domain loss functions for obtaining high fidelity source estimates. The approach offers superior flexibility because it is entirely an inference-time technique and can therefore efficiently deal with varying numbers of known sources in a given mixture, in contrast with standard supervised approaches which require re-training or extensive fine-tuning.
- The careful design of deep model priors that consider the data modality is critical for the task of inversion. In particular, it was identified that including adaptive dilated convolutions and dense connections to extract useful multi-scale features significantly improves the fidelity and computational complexity for ill-posed restoration tasks such as denoising, in-painting, and source separation. This design improves the robustness of the model to sampling rate changes and enables complex temporal modeling.

11.2 Future Work

The following comprise future research prospects that can potentially be an extension of this thesis.

- It was demonstrated that the DEP can effectively regularize the CF synthesis

process. Such a process can be benefited by enforcing the inversion to synthesize CFs with lower epistemic uncertainties by utilizing Δ-UQ. The differences in image synthesis in both methods can be studied quantitatively and qualitatively.

- An important aspect of improving model reliability is the ability to generalize to mild distribution shifts, commonly referred to as OOD generalization. This has been addressed by several approaches such as transfer learning in pre-trained models. Investigating how the principles of Δ-UQ can be used on pre-trained models to improve OOD generalization is an important research direction.
- Since DEPs serve as reliable failure detectors, the possibility of using them to predict generalization accuracy in deep models can be further explored. Additionally, there is a need to investigate ways of better leveraging or crafting the losses obtained to build improved OOD detectors.
- While latent space inlier augmentations and pixel space outlier augmentations were found to be critical for improving OOD detection, a theoretically justified framework can better support the understanding of the overall interactions taking place.
- In the problem of source separation with generative priors, the scenario where the mixing process and the sources were known *a priori* was considered. The problem of source separation can be further benefited by designing new strategies for handling cases where (i) the mixing process is unknown and (ii) the mixture contains novel sources. Replacing WaveGAN with improved audio generative models can perhaps benefit the overall inverse optimization.
- Rethinking source separation under the lens of self-supervised learning (Caron et al. 2021) is a potential direction. In this context, the selection of the pretext

task can be very critical when performing the downstream inference of source separation.

- While adaptive dilations and dense connections for constructing the layers of a deep model prior remain the state-of-the-art for restoration with CNN-based deep priors, the process can be benefited by adopting even more complex models such as audio transformers (Verma and Berger 2021) and modifying different layers within them to deal with the varying statistics of audio.
- With recent advancements in Quantum computing and Quantum Machine Learning (QML) in a variety of applications (Uehara, Spanias, and Clark 2021; G. Uehara et al. 2021; G. S. Uehara et al. 2022), it would be beneficial to extend the algorithms developed in this dissertation using QML. QML can offer extremely high data efficiency and computational benefits.

REFERENCES

Achanta, Radhakrishna, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. "SLIC superpixels compared to state-of-the-art superpixel methods." *IEEE transactions on pattern analysis and machine intelligence* 34 (11): 2274–2282.

AI Weekly. 2022. *AI Weekly: U.S. agencies are increasing their AI investments.* Https://venturebeat.com/ai/ai-weekly-u-s-agencies-are-increasing-their-investments-in-ai/, Last accessed on 2022-9-9.

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, et al. 2016. "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin." In *Proceedings of The 33rd International Conference on Machine Learning,* edited by Maria Florina Balcan and Kilian Q. Weinberger, 48:173–182. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June. https://proceedings.mlr.press/v48/amodei16.html.

Andreas, Kirsch, Mukhoti Jishnu, Joost van Amersfoort, Torr Philip H.S., and Gal Yarin. 2021. "On Pitfalls in OoD Detection: Entropy Considered Harmful." *ICML UDL Workshop.*

Anirudh, J.J, R.and Thiagarajan. 2021. "Δ-UQ: Accurate Uncertainty Quantification via Anchor Marginalization." *arXiv preprint arxiv:2110.02197.*

Anirudh, R, and J.J Thiagarajan. 2022. "Out of Distribution Detection via Neural Network Anchoring." In *Asian Conference on Machine Learning (ACML).* PMLR.

Anirudh, Rushil, Jayaraman J Thiagarajan, Bhavya Kailkhura, and Peer-Timo Bremer. 2020. "Mimicgan: Robust projection onto image manifolds with corruption mimicking." *International Journal of Computer Vision,* 1–19.

Arora, Sanjeev, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. 2019. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks." In *International Conference on Machine Learning,* 322–332. PMLR.

Ash, Jordan T, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds." In *International Conference on Learning Representations.*

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems* 33:12449–12460.

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. "A theory of learning from different domains." *Machine learning* 79 (1): 151–175.

Bietti, Alberto, and Julien Mairal. 2019. "On the inductive bias of neural tangent kernels." *Advances in Neural Information Processing Systems* 32.

Bishop, Christopher M, and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning.* Vol. 4. Springer.

Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. "Weight uncertainty in neural network." In *International Conference on Machine Learning,* 1613–1622. PMLR.

Bora, Ashish, Ajil Jalal, Eric Price, and Alexandros G Dimakis. 2017. "Compressed sensing using generative models." *34th International Conference on Machine Learning (ICML)* 70:537–546.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. "Language models are few-shot learners." *Advances in neural information processing systems* 33:1877–1901.

Cao, Tianshi, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. 2020. "A benchmark of medical out of distribution detection." *arXiv preprint arXiv:2007.04250.*

Cao, Tianshi, David Yu-Tung Hui, Chinwei Huang, and Joseph Paul Cohen. 2020. "Which MOoD Methods work? A Benchmark of Medical Out of Distribution Detection."

Caron, Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. "Emerging properties in self-supervised vision transformers." In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 9650–9660.

Chang, C., E. Creager, A. Goldenberg, and D. Duvenaud. 2018. "Explaining Image Classifiers by Counterfactual Generation." In *International Conference on Learning Representations.*

Chang, Ya-Liang, Kuan-Ying Lee, Po-Yu Wu, Hung-yi Lee, and Winston Hsu. 2019. "Deep long audio inpainting." *arXiv preprint arXiv:1911.06476.*

Chen, Jiefeng, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. 2021. "Atom: Robustifying out-of-distribution detection using outlier mining." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* 430–445. Springer.

Chen, Y., S. Liu, and X. Wang. 2021. "Learning continuous image representation with local implicit image function." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 8628–8638.

Codella, N., V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. 2019. "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)." *arXiv preprint arXiv:1902.03368.*

Codella, Noel CF, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 2018. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)." In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018),* 168–172.

Combalia, Marc, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. 2019. "BCN20000: Dermoscopic lesions in the wild." *arXiv preprint arXiv:1908.02288.*

Daras, G., J. Dean, A. Jalal, and AG. Dimakis. 2021. "Intermediate layer optimization for inverse problems using deep generative models." *arXiv preprint arXiv:2102.07364.*

Défossez, Alexandre, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. "Demucs: Deep extractor for music sources with extra unlabeled data remixed." *arXiv preprint arXiv:1909.01174.*

Défossez, Alexandre, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. 2018. "Sing: Symbol-to-instrument neural generator." *Advances in Neural Information Processing Systems,* 9041–9051.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. "ImageNet: A large-scale hierarchical image database." In *2009 IEEE Conference on Computer*

Vision and Pattern Recognition, 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Deng, Weijian, and Liang Zheng. 2021. "Are Labels Always Necessary for Classifier Accuracy Evaluation?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 15069–15078.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *NAACL-HLT (1),* 4171–4186. https://aclweb.org/anthology/papers/N/N19/N19-1423/.

Dhurandhar, A., P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das. 2018. "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." In *Advances in Neural Information Processing Systems,* vol. 31.

Donahue, Chris, Julian McAuley, and Miller Puckette. 2019. "Adversarial Audio Synthesis." *ICLR.*

Dosovitskiy, A., and T. Brox. 2016. "Inverting visual representations with convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4829–4837.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *International Conference on Learning Representations.* https://openreview.net/forum?id=YicbFdNTTy.

Du, Xuefeng, Zhaoning Wang, Mu Cai, and Yixuan Li. 2022. "VOS: Learning What You Don't Know by Virtual Outlier Synthesis." *arXiv preprint arXiv:2202.01197.*

Dua, Dheeru, and Casey Graff. 2017. *UCI Machine Learning Repository.* Http://archive.ics.uci.edu/ml, Last accessed on 08/01/2020. University of California, Irvine, School of Information and Computer Sciences.

Fei, Geli, and B. Liu. 2016. "Breaking the Closed World Assumption in Text Classification." In *NAACL.*

Févotte, Cédric, Emmanuel Vincent, and Alexey Ozerov. 2018. "Single-channel audio source separation with NMF: Divergences, constraints and algorithms." *Audio Source Separation,* 1–24.

Fort, Stanislav, Huiyi Hu, and Balaji Lakshminarayanan. 2019. "Deep ensembles: A loss landscape perspective." *arXiv preprint arXiv:1912.02757.*

G. Matthews, Alexander G. de, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. 2018. "Gaussian Process Behaviour in Wide Deep Neural Networks." In *International Conference on Learning Representations.* https://openreview.net/forum?id=H1-nGgWC-.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In *International Conference on Machine Learning,* 1050–1059.

Gandelsman, Yosef, Assaf Shocher, and Michal Irani. 2019. ""Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11026–11035.

Geoffrey, H., Oriol V., and Jeffrey D. 2015. "Distilling the Knowledge in a Neural Network." In *NIPS Deep Learning and Representation Learning Workshop.*

Gessert, Nils, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. 2020. "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data." *MethodsX* 7:100864.

Giri, Ritwik, Umut Isik, and Arvindh Krishnaswamy. 2019. "Attention wave-u-net for speech enhancement." In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* 249–253. IEEE.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. "Domain adaptation for large-scale sentiment classification: A deep learning approach." In *ICML.*

Gokhale, Tejas, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. 2022. "Improving Diversity with Adversarially Learned Transformations for Domain Generalization." *arXiv preprint arXiv:2206.07736.*

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative adversarial nets." In *Advances in neural information processing systems,* 2672–2680.

Goyal, Y., Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. 2019. "Counterfactual visual explanations." In *International Conference on Machine Learning,* 2376–2384. PMLR.

Grais, Emad M, Dominic Ward, and Mark D Plumbley. 2018. "Raw Multi-Channel Audio Source Separation using Multi-Resolution Convolutional Auto-Encoders." In *2018 26th European Signal Processing Conference (EUSIPCO),* 1577–1581. IEEE.

Granger, Clive WJ. 1969. "Investigating causal relations by econometric models and cross-spectral methods." *Econometrica: journal of the Econometric Society,* 424–438.

Graves, Alex. 2011. "Practical variational inference for neural networks." In *Advances in neural information processing systems,* 2348–2356. Citeseer.

Gretton, Arthur, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. "A Kernel Two-Sample Test." *Journal of Machine Learning Research* 13 (25): 723–773. http://jmlr.org/papers/v13/gretton12a.html.

Guillory, Devin, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. 2021. "Predicting with Confidence on Unseen Distributions." *arXiv preprint arXiv:2107.03315.*

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. "Improved training of wasserstein gans." *Advances in neural information processing systems,* 5767–5777.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. "On calibration of modern neural networks." In *International conference on machine learning,* 1321–1330. PMLR.

He, Bobby, Balaji Lakshminarayanan, and Yee Whye Teh. 2020. "Bayesian deep ensembles via the neural tangent kernel." *Advances in Neural Information Processing Systems* 33:1010–1022.

Hendrycks, D., M. Mazeika, and T. Dietterich. 2018. "Deep Anomaly Detection with Outlier Exposure." In *International Conference on Learning Representations.*

Hendrycks, Dan, and Kevin Gimpel. 2017. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks." *Proceedings of International Conference on Learning Representations.*

Hendrycks, Dan, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. 2020. "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty." *Proceedings of the International Conference on Learning Representations (ICLR).*

Hendrycks, Dan, Andy Zou, Mantas Mazeika, Leonard Tang, Dawn Song, and Jacob Steinhardt. 2021. "PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures." In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications.*

Hendrycks, Dan, and Thomas Dietterich. 2019. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations." In *International Conference on Learning Representations.* https://openreview.net/forum?id=HJz6tiCqYm.

Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. 2018. "Artificial intelligence in radiology." *Nature Reviews Cancer* 18 (8): 500–510.

Hsu, Y. -C., Y. Shen, H. Jin, and Z. Kira. 2020. "Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 10948–10957. https://doi.org/10.1109/CVPR42600.2020.01096.

Hull, J. J. 1994. "A database for handwritten text recognition research." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16 (5): 550–554.

Hyvärinen, Aapo. 1999. "Survey on independent component analysis." *Citeseer.*

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167.*

Ito, Keith, and Linda Johnson. 2017. *The LJ Speech Dataset.* https://keithito.com/LJ-Speech-Dataset/.

Jacot, Arthur, Franck Gabriel, and Clément Hongler. 2018. "Neural tangent kernel: Convergence and generalization in neural networks." *Advances in neural information processing systems* 31.

Jain, Moksh, Salem Lahlou, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. 2021. "DEUP: Direct Epistemic Uncertainty Prediction." *arXiv preprint arXiv:2102.08501.*

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. 2013. "Quantifying causal influences." *The Annals of Statistics* 41 (5): 2324–2358.

Jiang, Yiding, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. 2019. "Predicting the Generalization Gap in Deep Networks with Margin Distributions." In *International*

Conference on Learning Representations. https://openreview.net/forum?id=HJlQfnCqKX.

Kailkhura, Bhavya, Jayaraman J Thiagarajan, Charvi Rastogi, Pramod K Varshney, and Peer-Timo Bremer. 2018. "A spectral approach for the design of experiments: Design, analysis and algorithms." *The Journal of Machine Learning Research* 19 (1): 1214–1259.

Karhunen, Juha, Liuyue Wang, and Ricardo Vigario. 1995. "Nonlinear PCA type approaches for source separation and independent component analysis." *International Conference on Neural Networks (ICNN)* 2:995–1000.

Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. "Alias-free generative adversarial networks." *arXiv preprint arXiv:2106.12423.*

Karras, Tero, Samuli Laine, and Timo Aila. 2019. "A style-based generator architecture for generative adversarial networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4401–4410.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. "Analyzing and improving the image quality of stylegan." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 8110–8119.

Kather, Jakob Nikolas, Niels Halama, and Alexander Marx. 2018. "100,000 histological images of human colorectal cancer and healthy tissue." *Zenodo* (April). https://doi.org/10.5281/zenodo.1214456.

Kather, Jakob Nikolas, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. 2019. "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study." *PLoS medicine* 16 (1): e1002730.

Kingma, DP., and J. Ba. 2015. "Adam: A Method for Stochastic Optimization." *International Conference on Learning Representations, San Diego, California.*

Krishnan, Ranganath, and Omesh Tickoo. 2020. "Improving model calibration with accuracy versus uncertainty optimization." In *Advances in Neural Information Processing Systems,* 33:18237–18248.

Krizhevsky, Alex, and Geoffrey Hinton. 2009. "Learning multiple layers of features from tiny images." *Citeseer.*

Kuleshov, Volodymyr, Nathan Fenner, and Stefano Ermon. 2018. "Accurate uncertainties for deep learning using calibrated regression." In *International Conference on Machine Learning,* 2796–2804. PMLR.

Lakkaraju, Hima, Nino Arsov, and Osbert Bastani. 2020. "Robust Black Box Explanations Under Distribution Shift." *International Conference on Machine Learning (ICML).*

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. "Faithful and customizable explanations of black box models." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society,* 131–138.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. 2017. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30.

Lampinen, Jouko, and Aki Vehtari. 2001. "Bayesian approach for neural networks—review and case studies." *Neural networks* 14 (3): 257–274.

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

LeCun, Yann, Corinna Cortes, and CJ Burges. 2010. "MNIST handwritten digit database." *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2.

Lee, Jaehoon, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. 2018. "Deep Neural Networks as Gaussian Processes." In *International Conference on Learning Representations.* https://openreview.net/forum?id=B1EA-M-0Z.

Lee, Jaehoon, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. 2019. "Wide neural networks of any depth evolve as linear models under gradient descent." *Advances in neural information processing systems* 32.

Lee, Kimin, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." *Advances in neural information processing systems* 31.

Li, Da, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. 2017. "Deeper, broader and artier domain generalization." In *Proceedings of the IEEE international conference on computer vision,* 5542–5550.

Liang, Shiyu, Yixuan Li, and R Srikant. 2018. "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks." In *International Conference on Learning Representations.*

Liu, Z., P. Luo, X. Wang, and X. Tang. 2015. "Deep Learning Face Attributes in the Wild." In *Proceedings of International Conference on Computer Vision (ICCV).* December.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. "A ConvNet for the 2020s." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Lluis, Francesc, Jordi Pons, and Xavier Serra. 2018. "End-to-end music source separation: is it possible in the waveform domain?" *arXiv preprint arXiv:1810.12187.*

Looveren, AV., and J. Klaise. 2021. "Interpretable counterfactual explanations guided by prototypes." In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* 650–665. Springer.

Lundberg, Scott M, and Su-In Lee. 2017. "A unified approach to interpreting model predictions." In *Advances in neural information processing systems,* 4765–4774.

Luo, Yi, and Nima Mesgarani. 2019. "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (8): 1256–1266.

Mahendran, A., and A. Vedaldi. 2015. "Understanding deep image representations by inverting them." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 5188–5196.

Makino, S., S. Araki, R. Mukai, and H. Sawada. 2004. "Audio source separation based on independent component analysis." In *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No.04CH37512),* vol. 5, V–V, May, 2004. https://doi.org/10.1109/ISCAS.2004.1329896.

Michelashvili, Michael, and Lior Wolf. 2019. "Speech Denoising by Accumulating Per-Frequency Modeling Fluctuations." *arXiv e-prints,* arXiv–1904.

Mika, Sebastian, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. 1999. "Kernel PCA and de-noising in feature spaces." *Advances in neural information processing systems,* 536–542.

Mordvintsev, A., C. Olah, and M. Tyka. 2017. *Inceptionism: Going deeper into neural networks.* Https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, Last accessed on 2021-5-27.

Morgado, Pedro, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. "Self-supervised generation of spatial audio for 360 video." *Advances in Neural Information Processing Systems,* 362–372.

Müller, Samuel G., and Frank Hutter. 2021. "TrivialAugment: Tuning-Free Yet State-of-the-Art Data Augmentation." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 774–782. October.

Narayanaswamy, V., R. Ayyanar, C. Tepedelenlioglu, D. Srinivasan, and A. Spanias. 2023. "Optimizing Solar Power Using Array Topology Reconfiguration With Regularized Deep Neural Networks." *IEEE Access, Accepted.*

Narayanaswamy, Vivek, Rushil Anirudh, Irene Kim, Yamen Mubarka, Andreas Spanias, and Jayaraman J. Thiagarajan. 2022. "Predicting the Generalization Gap in Deep Models using Anchoring." In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 4393–4397. https://doi.org/10.1109/ICASSP43922.2022.9747136.

Narayanaswamy, Vivek, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, Andreas Spanias, and Jayaraman J Thiagarajan. 2022. "Improved Medical Out-of-Distribution Detectors For Modality and Semantic Shifts." *ICML Workshop on Principles of Distribution Shifts.*

Narayanaswamy, Vivek, Deepta Rajan, Andreas Spanias, and Jayaraman J Thiagarajan. 2022. "Using Direct Error Predictors to Improve Model Safety and Interpretability." *ICML Workshop on Interpretable Machine Learning in Healthcare.*

Narayanaswamy, Vivek, Rakshith Subramanyam, Mark Naufel, Andreas Spanias, and Jayaraman J. Thiagarajan. 2022. "Improved StyleGAN-v2 based Inversion for Out-of-Distribution Images." In *Proceedings of the 39th International Conference on Machine Learning,* edited by Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, 162:20625–20639. Proceedings of Machine Learning Research. PMLR, July. https://proceedings.mlr.press/v162/subramanyam22a.html.

Narayanaswamy, Vivek, Jayaraman J Thiagarajan, Rushil Anirudh, and Andreas Spanias. 2020. "Unsupervised audio source separation using generative priors." In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH,* 2020:2657–2661.

Narayanaswamy, Vivek, Jayaraman J Thiagarajan, Deepta Rajan, and Andreas Spanias. 2021. "Loss Estimators Improve Model Generalization." *arXiv preprint arXiv:2103.03788.*

Narayanaswamy, Vivek, Jayaraman J Thiagarajan, and Andreas Spanias. 2021. "Using deep image priors to generate counterfactual explanations." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2770–2774. IEEE.

Narayanaswamy, Vivek Sivaraman, Sameeksha Katoch, Jayaraman J Thiagarajan, Huan Song, and Andreas Spanias. 2019. "Audio Source Separation via Multi-Scale Learning with Dilated Dense U-Nets." *arXiv preprint arXiv:1904.04161.*

Narayanaswamy, Vivek Sivaraman, Jayaraman J Thiagarajan, Huan Song, and Andreas Spanias. 2019. "Designing an effective metric learning pipeline for speaker diarization." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 5806–5810. IEEE.

Narayanaswamy, Vivek Sivaraman, Jayaraman J Thiagarajan, and Andreas Spanias. 2021. "On the design of deep priors for unsupervised audio restoration." In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021,* 2858–2862. International Speech Communication Association.

Neal, Radford M. 2012. *Bayesian learning for neural networks.* Vol. 118. Springer Science & Business Media.

Nemirovsky, D., N. Thiebaut, Y. Xu, and A. Gupta. 2020. "CounteRGAN: Generating Realistic Counterfactuals with Residual Generative Adversarial Nets." *arXiv preprint arXiv:2009.05199.*

Novak, Roman, Lechao Xiao, Yasaman Bahri, Jaehoon Lee, Greg Yang, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-dickstein. 2019. "Bayesian Deep Convolutional Networks with Many Channels are Gaussian Processes." In *International Conference on Learning Representations.* https://openreview.net/forum?id=B1g30j0qF7.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499.*

Ovadia, Yaniv, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift." *Advances in neural information processing systems* 32.

Pacheco, Andre GC, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. 2020. "On Out-of-Distribution Detection Algorithms with Deep Neural Skin Cancer Classifiers." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops,* 732–733.

Pascual, Santiago, Antonio Bonafonte, and Joan Serra. 2017. "SEGAN: Speech enhancement generative adversarial network." *arXiv preprint arXiv:1703.09452.*

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Radford, Alec, Luke Metz, and Soumith Chintala. 2015. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434.*

Rao, Qing, and Jelena Frtunikj. 2018. "Deep Learning for Self-Driving Cars: Chances and Challenges." In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems,* 35–38. SEFAIS '18. Gothenburg, Sweden: Association for Computing Machinery. https://doi.org/10.1145/3194085.3194087.

Rao, Sunil, Sameeksha Katoch, Vivek Narayanaswamy, Gowtham Muniraju, Cihan Tepedelenlioglu, Andreas Spanias, Pavan Turaga, Raja Ayyanar, and Devarajan Srinivasan. 2020. "Machine learning for solar array monitoring, optimization, and control." *Synthesis Lectures on Power Electronics* 7 (1): 1–91.

Rao, Sunil, Vivek Narayanaswamy, Michael Esposito, Jayaraman Thiagarajan, and Andreas Spanias. 2021. "Deep Learning with hyper-parameter tuning for COVID-19 Cough Detection." In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA),* 1–5. IEEE.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. "Do imagenet classifiers generalize to imagenet?" In *International Conference on Machine Learning,* 5389–5400. PMLR.

Ren, Jie, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. 2021. "A simple fix to mahalanobis distance for improving near-ood detection." *arXiv preprint arXiv:2106.09022.*

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

Ribeiro, M.T., S. Singh, and C. Guestrin. 2016. "" Why should I trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* 1135–1144.

Liu, Weitang, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. "Energy-based Out-of-distribution Detection." *Advances in Neural Information Processing Systems.*

Mahendran, A., and A. Vedaldi. 2016. "Visualizing deep convolutional neural networks using natural pre-images." *International Journal of Computer Vision* 120 (3): 233–255.

Narayanaswamy, Vivek, Yamen Mubarka, Rushil Anirudh, Deepta Rajan, Andreas Spanias, and Jayaraman J Thiagarajan. 2022. "Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection." *arXiv preprint arXiv:2207.05286.*

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

Roy, Abhijit Guha, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. 2022. "Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions." *Medical Image Analysis* 75:102274.

Ribeiro, MT., S. Singh, and C. Guestrin. 2018. "Anchors: High-precision model-agnostic explanations." In *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al.

2015. “Imagenet large scale visual recognition challenge.” *International journal of computer vision* 115 (3): 211–252.

Virtanen, Tuomas. 2007. “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria.” *IEEE transactions on audio, speech, and language processing* 15 (3): 1066–1074.

Sastry, Chandramouli Shama, and Sageev Oore. 2020. “Detecting out-of-distribution examples with gram matrices.” In *International Conference on Machine Learning,* 8491–8501. PMLR.

Sauer, A., and A. Geiger. 2021. “Counterfactual Generative Networks.” *ICLR.*

Schwab, Patrick, and Helmut Hlavacs. 2015. “Capturing the Essence: Towards the Automated Generation of Transparent Behavior Models.” In *AIIDE,* 184–190.

Schwab, Patrick, and Walter Karlen. 2019. “CXPlain: Causal explanations for model interpretation under uncertainty.” In *Advances in Neural Information Processing Systems,* 10220–10230.

Schwab, Patrick, Djordje Miladinovic, and Walter Karlen. 2019. “Granger-causal attentive mixtures of experts: Learning important features with neural networks.” In *Proceedings of the AAAI Conference on Artificial Intelligence,* 33:4846–4853.

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

Seo, Seonguk, Paul Hongsuck Seo, and Bohyung Han. 2019. “Learning for Single-Shot Confidence Calibration in Deep Neural Networks through Stochastic Inferences.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 9030–9038.

Shah, Neil, Nandish Bhagat, and Manan Shah. 2021. “Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention.” *Visual Computing for Industry, Biomedicine, and Art* 4 (1): 1–14.

Shah, Viraj, and Chinmay Hegde. 2018. “Solving linear inverse problems using gan priors: An algorithm with provable guarantees.” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 4609–4613.

Shahriari, Bobak, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. 2015. "Taking the human out of the loop: A review of Bayesian optimization." *Proceedings of the IEEE* 104 (1): 148–175.

Shanthamallu, Uday Shankar, and Andreas Spanias. 2022. "Machine and Deep Learning Applications." *Machine and Deep Learning Algorithms and Applications* (Cham), 59–72.

Shanthamallu, Uday Shankar, Andreas Spanias, Cihan Tepedelenlioglu, and Mike Stanley. 2017. "A brief survey of machine learning methods and their sensor and IoT applications." In *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA),* 1–8. https://doi.org/10.1109/IISA.2017.8316459.

Shorten, Connor, and Taghi M Khoshgoftaar. 2019. "A survey on image data augmentation for deep learning." *Journal of big data* 6 (1): 1–48.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. "Learning Important Features Through Propagating Activation Differences." In *Proceedings of Machine Learning Research,* 70:3145–3153.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *arXiv preprint arXiv:1312.6034.*

Sinha, Abhishek, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. 2021. "Negative Data Augmentation." In *International Conference on Learning Representations.*

Sitzmann, V., J. Martel, A. Bergman, D. Lindell, and G. Wetzstein. 2020. "Implicit neural representations with periodic activation functions." *Advances in Neural Information Processing Systems* 33.

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams. 2012. "Practical bayesian optimization of machine learning algorithms." *Advances in neural information processing systems* 25.

Spanias, Andreas. July 2015. "Advances in speech and audio processing and coding." *6th IEEE International Conference on Information, Intelligence, Systems and Applications (IISA),* 1–2.

Spanias, Andreas, Ted Painter, and Venkatraman Atti. 2006. *Audio signal processing and coding.* John Wiley & Sons.

Spiertz, Martin, and Volker Gnann. 2009. "Source-filter based clustering for monaural blind source separation." *Proceedings of the 12th International Conference on Digital Audio Effects.*

Stoller, Daniel, Sebastian Ewert, and Simon Dixon. 2018. "Wave-u-net: A scale neural network for end-to-end audio source separation." *arXiv preprint arXiv:1806.03185.*

Stöter, Fabian-Robert, Antoine Liutkus, and Nobutaka Ito. 2018. "The 2018 Signal Separation Evaluation Campaign." In *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK,* 293–305.

Štrumbelj, Erik, Igor Kononenko, and M Robnik Šikonja. 2009. "Explaining instance classifications with interactions of subsets of feature values." *Data & Knowledge Engineering* 68 (10): 886–904.

Subramanyam, Rakshith, Mark Heimann, Jayram Thathachar, Rushil Anirudh, and Jayaraman J Thiagarajan. 2022. "Contrastive Knowledge-Augmented Meta-Learning for Few-Shot Classification." *arXiv preprint arXiv:2207.12346.*

Sumedha, S., P. Brian, C. Junxiang, and B. Kayhan. 2020. "Explanation by Progressive Exaggeration." In *International Conference on Learning Representations.*

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. 2017. "Axiomatic attribution for deep networks." *arXiv preprint arXiv:1703.01365.*

Takahashi, Naoya, Nabarun Goswami, and Yuki Mitsufuji. 2018. "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation." In *2018 16th IEEE International Workshop on Acoustic Signal Enhancement (IWAENC),* 106–110, 2018.

Takeuchi, Hirotoshi, Kunio Kashino, Yasunori Ohishi, and Hiroshi Saruwatari. 2020. "Harmonic Lowering for Accelerating Harmonic Convolution for Audio Signals." *Proc. Interspeech 2020,* 185–189.

Tancik, M., PP. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, JT. Barron, and R. Ng. 2020. "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains." *NeurIPS.*

Thiagarajan, J.J, P. Sattigeri, D. Rajan, and B. Venkatesh. 2020. "Calibrating Healthcare AI: Towards Reliable and Interpretable Deep Predictive Models." *arXiv preprint arXiv:2004.14480.*

Thiagarajan, Jayaraman, Vivek Sivaraman Narayanaswamy, Deepta Rajan, Jia Liang, Akshay Chaudhari, and Andreas Spanias. 2021. "Designing counterfactual generators using deep model inversion." *Advances in Neural Information Processing Systems* 34:16873–16884.

Thiagarajan, Jayaraman J, Rushil Anirudh, Vivek Narayanaswamy, and Peer-Timo Bremer. 2022. "Single Model Uncertainty Estimation via Stochastic Data Centering." *Advances in Neural Information Processing Systems.*

Thiagarajan, Jayaraman J, Deepta Rajan, Sameeksha Katoch, and Andreas Spanias. 2020. "DDxNet: a deep learning model for automatic interpretation of electronic health records, electrocardiograms and electroencephalograms." *Scientific reports* 10 (1): 1–11.

Thiagarajan, Jayaraman J, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. 2013. "Mixing matrix estimation using discriminative clustering for blind source separation." *Digital Signal Processing* 23 (1): 9–18.

Thiagarajan, Jayaraman J, Bindya Venkatesh, and Deepta Rajan. 2020. "Learn-by-calibrating: Using calibration as a training objective." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 3632–3636. IEEE.

Thiagarajan, Jayaraman J, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer. 2020. "Building Calibrated Deep Models via Uncertainty Matching with Auxiliary Interval Predictors." In *AAAI,* 6005–6012.

Thiagarajan, JJ., V. Narayanaswamy, R. Anirudh, PT. Bremer, and A. Spanias. 2021. "Accurate and Robust Feature Importance Estimation under Distribution Shifts." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 35:7891–7898.

Thiemann, J, N Ito, and E Vincent. 2013. *Diverse Environments Multichannel Acoustic Noise Database (DEMAND).*

Thulasidasan, Sunil, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. 2021. *A Simple and Effective Baseline for Out-of-Distribution Detection using Abstention.* https://openreview.net/forum?id=q_Q9MMGwSQu.

Tian, Yapeng, Chenliang Xu, and Dingzeyu Li. 2019. "Deep Audio Prior." *arXiv preprint arXiv:1912.10292.*

Tran, Dustin, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. 2022. "Plex: Towards reliability using pretrained large model extensions." *arXiv preprint arXiv:2207.07411.*

Tschandl, P., C. Rosendahl, and H. Kittler. 2018. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions." *Scientific data* 5:180161.

Uehara, Glen, Sunil Rao, Mathew Dobson, Cihan Tepedelenlioglu, and Andreas Spanias. 2021. "Quantum Neural Network Parameter Estimation for Photovoltaic Fault Detection." In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA),* 1–7. IEEE.

Uehara, Glen S, Vivek Narayanaswamy, Cihan Tepedelenlioglu, and Andreas Spanias. 2022. "Quantum machine learning for photovoltaic topology optimization." In *2022 13th International Conference on Information, Intelligence, Systems & Applications (IISA),* 1–5. IEEE.

Uehara, Glen S, Andreas Spanias, and William Clark. 2021. "Quantum information processing algorithms with emphasis on machine learning." In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA),* 1–11. IEEE.

Ulyanov, Dmitry, Andrea Vedaldi, and Victor Lempitsky. 2018. "Deep image prior." *IEEE Conference on Computer Vision and Pattern Recognition,* 9446–9454.

Van Amersfoort, J., L. Smith, YW. Teh, and Y. Gal. 2020. "Uncertainty estimation using a single deep deterministic neural network." In *International Conference on Machine Learning,* 9690–9700. PMLR.

Vanschoren, Joaquin, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. "OpenML: Networked Science in Machine Learning." *SIGKDD Explorations* (New York, NY, USA) 15 (2): 49–60. https://doi.org/10.1145/2641190.2641198.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Advances in Neural Information Processing Systems,* 6000–6010.

Verma, Prateek, and Jonathan Berger. 2021. "Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions." *arXiv preprint arXiv:2105.00335.*

Verma, S., J. Dickerson, and K. Hines. 2020. "Counterfactual Explanations for Machine Learning: A Review." *arXiv preprint arXiv:2010.10596.*

Virtanen, Tuomas. 2003. "Sound Source Separation Using Sparse Coding with Temporal Continuity Objective." *ICMC,* 231–234.

Wang, Haotao, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. 2021. "AugMax: Adversarial Composition of Random Augmentations for Robust Training." In *NeurIPS.*

Wang, Jason, Luis Perez, et al. 2017. "The effectiveness of data augmentation in image classification using deep learning." *Convolutional Neural Networks Vis. Recognit* 11:1–8.

Wang, Lin, Joshua D Reiss, and Andrea Cavallaro. 2016. "Over-determined source separation and localization using distributed microphones." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24 (9): 1573–1588.

Warden, Pete. 2018. "Speech commands: A dataset for limited-vocabulary speech recognition." *arXiv preprint arXiv:1804.03209.*

Welling, Max, and Yee W Teh. 2011. "Bayesian learning via stochastic gradient Langevin dynamics." In *Proceedings of the 28th international conference on machine learning (ICML-11),* 681–688. Citeseer.

Wilson, Andrew G, and Pavel Izmailov. 2020. "Bayesian deep learning and a probabilistic perspective of generalization." *Advances in neural information processing systems* 33:4697–4708.

Woodbury, Max A. 1950. *Inverting modified matrices.* Statistical Research Group.

Xu, Zhenlin, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. 2021. "Robust and Generalizable Visual Representation Learning via Random Convolutions." In *International Conference on Learning Representations.*

Yang, Jingkang, Kaiyang Zhou, and Ziwei Liu. 2022. "Full-Spectrum Out-of-Distribution Detection." *arXiv preprint arXiv:2204.05306.*

Yin, H., P. Molchanov, JM. Alvarez, Z. Li, A. Mallya, D. Hoiem, NK. Jha, and J. Kautz. 2020. "Dreaming to distill: Data-free knowledge transfer via deepinversion." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 8715–8724.

Yoo, Donggeun, and In So Kweon. 2019. "Learning loss for active learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 93–102.

Young, Albert T, Mulin Xiong, Jacob Pfau, Michael J Keiser, and Maria L Wei. 2020. "Artificial intelligence in dermatology: a primer." *Journal of Investigative Dermatology* 140 (8): 1504–1512.

Yu, Fisher, and Vladlen Koltun. 2015. "Multi-scale context aggregation by dilated convolutions." *arXiv preprint arXiv:1511.07122.*

Zagoruyko, Sergey, and Nikos Komodakis. 2016. "Wide residual networks." *arXiv preprint arXiv:1605.07146.*

Zhang, Hongyi, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. "mixup: Beyond Empirical Risk Minimization." In *International Conference on Learning Representations.*

Zhang, Jingyang, Nathan Inkawhich, Yiran Chen, and Hai Li. 2021. "Fine-grained Out-of-Distribution Detection with Mixup Outlier Exposure." *arXiv preprint arXiv:2106.03917.*

Zhang, Xuaner, Ren Ng, and Qifeng Chen. 2018. "Single image reflection separation with perceptual losses." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 4786–4794.

Zhang, Yujia, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. ""Why Should You Trust My Explanation" Understanding Uncertainty in LIME Explanations." *arXiv preprint arXiv:1904.12991.*

Zhang, Zhoutong, Yunyun Wang, Chuang Gan, Jiajun Wu, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. 2019. "Deep audio priors emerge from harmonic convolutional networks." In *International Conference on Learning Representations.*

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision,* 2223–2232.