University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Bartłomiej Sadlej
Student no. 429589

Title in English

Bachelor's thesis
in MACHINE LEARNING

Supervisor:
prof. Przemysław Biecek
Wydział Matematyki Informatyki i Mechaniki

Warsaw, May 2017

Abstract

Understanding the decision-making process of image classifiers is an active area of research in machine learning. Current industry-standard methods focus on finding human-interpretable concepts or features that influence predictions in known data samples. However, we argue that this approach is limited in its ability to provide a comprehensive understanding of model behavior due to vast unexplored regions of the data manifold which can potentially lead to the same predictions. This work is a continuation of pioneering work on Generative Explainable AI (XAI) Sobieski et al. [2024] published at ICLR 2025 and introduces a novel framework of Invariant Images - sets of images that yield identical predictions under a given objective function, thereby establishing an equivalence relation. Our method EquiDiff allows for sampling meaningful and diverse high quality realistic examples from these invariant sets and as a result opens new possibilities for evaluating existing explainability methods on unknown data - simulating the real world.

Keywords

blabaliza różnicowa, fetory σ-ρ, fooizm, blarbarucja, blaba, fetoryka, baleronik

Thesis domain (Socrates-Erasmus subject area codes)

11.4 Sztuczna inteligencja

Subject classification

D. Software
D.127. Blabalgorithms
D.127.6. Numerical blabalysis

Tytuł pracy w języku polskim

Tytuł po polsku

Contents

Chapter 1

Introduction

The remarkable success of deep neural networks in computer vision has been accompanied by an equally pressing need to understand their decision-making processes. As these models are deployed in critical applications ranging from medical diagnosis to autonomous driving, the ability to explain and interpret their behavior becomes paramount for building trust, ensuring fairness, and identifying potential failure modes.

Current explainable AI (XAI) methods have made significant strides in providing insights into model behavior through various approaches including saliency maps Simonyan et al. [2014], concept activation vectors Kim et al. [2018], and gradient-based attribution methods Sundararajan et al. [2017]. However, these approaches share a fundamental limitation: they primarily operate within the confines of known training data or slight perturbations thereof, leaving vast regions of the input manifold unexplored.

1.1. Motivation and Problem Statement

Consider a trained image classifier that correctly identifies both a standard photograph of a dog and a highly stylized artistic rendering of the same animal. Traditional XAI methods would analyze these two specific instances, potentially identifying common features like shape or texture patterns. However, they would miss the broader question: what other visual representations would this model also classify as a dog with the same confidence?

This question is not merely academic. Understanding the full scope of inputs that lead to identical model predictions is crucial for several reasons:

1. **Robustness Assessment**: Identifying the complete set of equivalent inputs reveals potential vulnerabilities and edge cases that might not be present in training data.

2. **Bias Detection**: Systematic patterns within invariant sets can reveal spurious correlations and biases that the model has learned.

3. **Fairness Evaluation**: Understanding what variations preserve predictions helps assess whether models make decisions based on relevant features rather than protected attributes.

4. **Generalization Understanding**: The structure of invariant sets provides insights into how models generalize beyond their training distribution.

1.2. Our Approach: Generative Explainable AI

This thesis introduces a paradigm shift from traditional interpolative XAI methods to a generative approach. Instead of analyzing existing data points, we propose to synthesize new, meaningful examples that preserve model predictions, thereby exploring the *Invariant Set* – the complete collection of inputs that yield identical outputs under a given objective function.

Our method, EquiDiff (Equivariant Diffusion Sampling), combines score-based generative models with classifier guidance to sample high-quality, diverse images from these invariant sets. By leveraging the powerful generative capabilities of diffusion models, we can explore regions of the input space that may never have been encountered during training, providing a more comprehensive understanding of model behavior.

Figure 1.1 illustrates the conceptual distinction between our approach and current XAI methods. While traditional methods focus on explaining decisions within known data boundaries, generative XAI explores the broader space of possible inputs that lead to the same predictions.

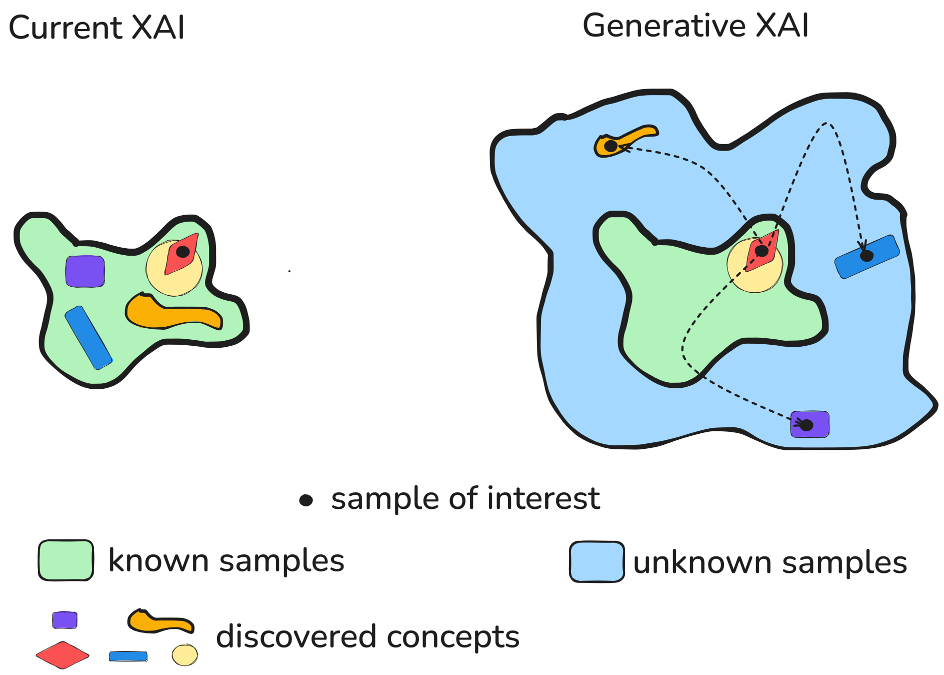

Figure 1.1: Conceptual comparison between traditional XAI methods and Generative XAI. Traditional methods analyze known data samples (left), while our approach synthesizes diverse examples from the invariant set that yield identical predictions (right).

1.3. Contributions

This thesis makes the following key contributions to the field of explainable AI:

1. **Theoretical Framework**: We provide a formal mathematical definition of invariant sets for neural network explanation and establish their properties as equivalence relations.
2. **Algorithmic Innovation**: We develop EquiDiff, a practical algorithm that combines score-based diffusion models with guided sampling to generate high-quality examples from invariant sets.

3. **Comprehensive Evaluation**: We conduct extensive experiments demonstrating the effectiveness of our approach across multiple computer vision tasks and architectures.

4. **Novel Insights**: We show how invariant set analysis can reveal model biases, spurious correlations, and learned invariances that are not detectable through traditional XAI methods.

5. **Quality Assurance**: We develop rigorous methods for ensuring the generated images maintain both semantic quality and mathematical invariance properties.

1.4. Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews relevant literature in explainable AI, generative modeling, and diffusion models. Chapter 3 presents our theoretical framework and the EquiDiff algorithm. Chapter 4 details our experimental setup and results. Chapter 5 explores practical applications of our framework. Chapter 6 discusses implications, limitations, and future directions. Finally, Chapter 7 summarizes our contributions and concludes the thesis.

Chapter 2

Related Work

This chapter reviews the relevant literature across several interconnected areas that form the foundation of our work. We begin with an overview of explainable AI methods, followed by background on score-based generative models, conditional generation techniques, and related work on activation maximization and concept discovery.

2.1. Explainable Artificial Intelligence

The field of explainable AI has evolved rapidly in response to the growing complexity and opacity of modern deep learning models. Current approaches can be broadly categorized into several paradigms:

2.1.1. Attribution Methods

Attribution methods aim to identify which input features are most important for a model's prediction. Gradient-based methods like Integrated Gradients Sundararajan et al. [2017] and GradCAM Selvaraju et al. [2017] compute the gradient of the output with respect to input features to determine importance scores. While computationally efficient, these methods are limited to local explanations around specific data points and can be sensitive to model architecture and input preprocessing.

Perturbation-based methods such as LIME Ribeiro et al. [2016] and SHAP Lundberg and Lee [2017] evaluate feature importance by measuring how predictions change when features are masked or altered. These methods provide more model-agnostic explanations but are computationally expensive and may not capture complex feature interactions.

2.1.2. Concept-Based Methods

Concept-based explainability methods attempt to understand models in terms of human-interpretable concepts. Concept Activation Vectors (CAVs) Kim et al. [2018] learn linear directions in activation space that correspond to human-defined concepts. Network Dissection Bau et al. [2017] automatically discovers concepts by correlating individual neurons with semantic segmentation labels.

More recent work has focused on discovering concepts automatically without human supervision. ACE (Automatic Concept Extraction) Ghorbani et al. [2019] uses unsupervised segmentation to identify important concepts, while TCAV (Testing with CAVs) Kim et al. [2018] provides statistical significance testing for concept importance.

2.1.3. Counterfactual Explanations

Counterfactual explanations answer the question "What would need to change for the model to make a different prediction?" This paradigm has gained popularity due to its intuitive nature and practical utility. Recent work by Sobieski et al. [2024] has pioneered the use of generative models for creating visual counterfactual explanations, directly inspiring our approach.

However, counterfactual methods typically focus on finding minimal changes that flip predictions, which is fundamentally different from our goal of finding diverse examples that preserve predictions.

2.2. Score-Based Generative Models

Score-based generative models (SGMs) have emerged as a powerful framework for high-quality image generation. Following the seminal work of Song et al. [2021], these models can be understood through the lens of stochastic differential equations (SDEs).

2.2.1. Mathematical Foundation

The core idea behind SGMs is to transform samples from a complex data distribution p_0 (e.g., natural images) to a simple noise distribution p_1 (typically Gaussian) through a forward diffusion process, then learn to reverse this transformation. The forward SDE is given by:

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t \tag{2.1}$$

where $\mathbf{x}_t$ represents the noisy version of a clean image at time $t \in [0, 1]$, $\mathbf{f}(\mathbf{x}_t, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and $\mathbf{w}_t$ is a Wiener process.

The corresponding reverse SDE, which enables generation, is:

$$\mathrm{d}\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g(t)^2\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t \tag{2.2}$$

The key term $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the score function, which must be learned by a neural network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$.

2.2.2. Training and Sampling

Score networks are typically trained using denoising score matching Vincent [2011]:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\lambda(t)\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_2^2\right] \tag{2.3}$$

where $\mathbf{x}_t = \alpha(t)\mathbf{x}_0 + \sigma(t)\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, and $\lambda(t)$ is a weighting function.

During sampling, we start from pure noise $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$ and integrate the reverse SDE using numerical solvers, with the learned score function $\mathbf{s}_{\boldsymbol{\theta}}$ approximating the true score.

2.3. Conditional Generation and Classifier Guidance

Conditional generation extends SGMs to produce samples conditioned on additional information $\mathbf{y}$, such as class labels or other attributes. The conditional score function can be decomposed using Bayes' theorem:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{y}, t) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t, t) \tag{2.4}$$

2.3.1. Classifier Guidance

Classifier guidance Dhariwal and Nichol [2021] implements conditional generation by training an auxiliary time-dependent classifier $p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t)$ on noisy images and incorporating its gradients into the sampling process:

$$\tilde{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{y}) = \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + s \cdot \nabla_{\mathbf{x}_t} \log p_{\boldsymbol{\phi}}(\mathbf{y} \mid \mathbf{x}_t, t) \tag{2.5}$$

where s is the guidance scale that controls the trade-off between sample quality and diversity.

2.3.2. Limitations of Standard Classifier Guidance

While effective for class-conditional generation, standard classifier guidance has several limitations for our application:

1. **Limited Optimization Steps**: The guidance is applied only during a fixed number of diffusion steps, which may be insufficient for precise conditioning.

2. **Latent Space Challenges**: Most state-of-the-art diffusion models operate in latent space, requiring careful handling of the conditioning signal.

3. **Objective Function Flexibility**: Standard methods are designed for classification tasks and may not easily extend to arbitrary objective functions.

These limitations motivate our approach, which we detail in Chapter 3.

2.4. Inverse Problems and Posterior Sampling

Recent work has explored the use of diffusion models as priors for solving inverse problems in image restoration Song et al. [2023], Chung et al. [2024]. The general inverse problem can be formulated as:

$$\mathbf{y} = \mathcal{A}(\mathbf{x}) + \boldsymbol{\epsilon} \tag{2.6}$$

where $\mathcal{A}$ is a (possibly nonlinear) forward operator, $\mathbf{x}$ is the unknown signal, $\mathbf{y}$ is the observed measurement, and $\boldsymbol{\epsilon}$ is noise.

Chung et al. [2024] showed that diffusion models can handle nonlinear inverse problems by incorporating the measurement likelihood into the reverse SDE. This work is particularly relevant to our approach, as neural network predictions can be viewed as nonlinear measurements of the input image.

2.4.1. Diverse Posterior Sampling

More recently, Cohen et al. [2024] extended inverse problem solvers to generate diverse solutions rather than a single best estimate. This paradigm shift from point estimation to posterior sampling aligns closely with our goal of generating diverse examples from invariant sets.

2.5. Activation Maximization and Feature Visualization

Activation maximization techniques attempt to synthesize inputs that maximally activate specific neurons or model outputs Erhan et al. [2009], Mordvintsev et al. [2015]. The basic approach optimizes an input image $\mathbf{x}$ to maximize an objective function $\mathcal{L}(\mathbf{x})$:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \mathcal{L}(\mathbf{x}) - \lambda \mathcal{R}(\mathbf{x}) \tag{2.7}$$

where $\mathcal{R}(\mathbf{x})$ is a regularization term to encourage natural-looking images.

However, activation maximization methods often produce unrealistic images with high-frequency artifacts that are imperceptible to humans but strongly activate neurons. Various regularization techniques have been proposed, including total variation penalties Mahendran and Vedaldi [2014] and frequency domain constraints Olah et al. [2017].

2.5.1. Limitations and Relationship to Our Work

While activation maximization shares the goal of understanding model behavior through synthetic inputs, it differs fundamentally from our approach:

1. **Single Solution vs. Diverse Sets**: Activation maximization typically finds one optimal input, while we aim to sample from the entire invariant set.
2. **Maximum Activation vs. Preserved Predictions**: Activation maximization seeks to maximize responses, while we preserve specific prediction values.
3. **Quality Issues**: Traditional activation maximization often produces unrealistic images, while our diffusion-based approach leverages strong visual priors.

2.6. Concept Discovery and Spurious Feature Detection

Understanding what concepts neural networks learn has been an active area of research. Lapuschkin et al. [2019] developed SpRAy, an automatic pipeline for exploring shortcuts and biases learned by models, often referred to as "Clever Hans" effects Pfungst [1911]. Neuhaus et al. [2023] investigates methods for automatically finding spurious features in training data.

Recent work by Dreyer et al. [2025] addresses the question of what concepts were learned by models and where in the training data they were present. However, Leask et al. [2025] argues that automatically discovered concepts may lack atomicity and completeness.

Our work complements this line of research by exploring the space of inputs that preserve predictions, potentially revealing spurious correlations and biases that may not be apparent from training data analysis alone.

2.7. Detection of Synthetic Images

As generative models become increasingly sophisticated, detecting synthetic images has become an important research area. Modern architectures using resampling operations (upsampling, downsampling, interpolation) introduce specific periodic correlations between pixels that are rarely present in natural images Popescu and Farid [2005].

Recent advances in synthetic image detection Zhang et al. [2019], Wang et al. [2023], Zhang and Xu [2023] have achieved near-perfect accuracy on images generated by GANs and diffusion models by analyzing frequency domain artifacts.

While we do not aim to fool detection methods (we acknowledge our images are synthetic), we incorporate insights from this literature to ensure our generated images encode meaningful signal in perceptually relevant frequency bands rather than imperceptible high-frequency artifacts.

2.8. Gap in Current Literature

Despite significant advances in explainable AI, a fundamental gap remains: current methods primarily analyze known training data and model behavior on observed inputs. This leaves vast regions of the input manifold unexplored, potentially missing important insights about model behavior.

Our work addresses this gap by introducing a principled framework for exploring the space of alternative inputs that yield identical predictions. By leveraging powerful generative models, we can sample from regions of the input space that may never have been encountered during training, providing a more comprehensive understanding of model behavior.

Chapter 3

Method

This chapter presents our theoretical framework and algorithmic approach for generating invariant sets. We begin with formal definitions, establish our theoretical foundation, detail our algorithm, and conclude with implementation specifics and quality assurance measures.

3.1. Problem Formulation

We formulate the problem of finding invariant sets (IS) as discovering members of an equivalence relation. Given an objective function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ and a query point $\mathbf{x}^*$, we define the invariant set as:

$$\mathbf{IS}(\mathbf{x}^*) = \{\mathbf{x} \in \mathbb{R}^n : \mathcal{L}(\mathbf{x}) = \mathcal{L}(\mathbf{x}^*)\} \tag{3.1}$$

We use the notation $\mathbf{x}^* \sim_{\mathcal{L}} \mathbf{x}$ to denote that two elements $\mathbf{x}^*$ and $\mathbf{x}$ belong to the same invariant set under the equivalence relation defined by $\mathcal{L}$.

The objective function $\mathcal{L}$ can represent various neural network components: a single neuron's activation, class logits for one or multiple classes, or any differentiable function for which gradients can be computed. While established methods exist for generating adversarial examples directly from $\mathbf{x}^*$ Szegedy et al. [2014], our goal is to sample meaningful and diverse examples from the entire invariant set.

To achieve this diversity and realism, we utilize a trained diffusion model, specifically LightningDIT Yao et al. [2025] Yao et al. [2024], which excels at generating high-quality images while maintaining the mathematical constraints of invariant set membership.

3.2. Guided Infinite Optimization with Latent Diffusion Models

Our algorithm integrates signals from the objective function $\mathcal{L} : \mathbb{R}^{W \times H} \to \mathbb{R}^m$ to conditionally synthesize images from invariant sets. There are two primary approaches for conditioning generation using $\mathcal{L}$.

3.2.1. Classifier Guidance Limitations

Classifier Guidance (CG) Dhariwal and Nichol [2021] offers a simple, computationally efficient method for trading diversity for fidelity using gradients from the objective function at each denoising step. However, we identified two significant limitations:

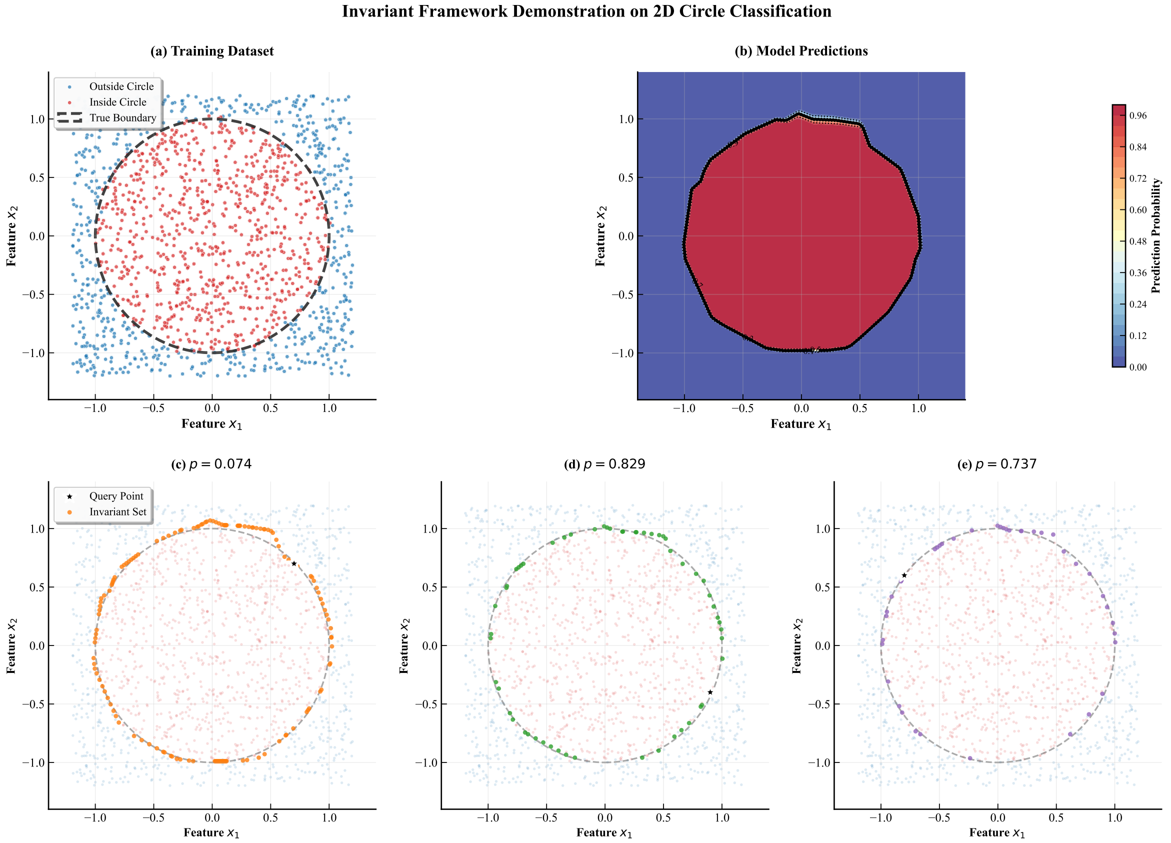

Figure 3.1: Demonstration of the Invariant Framework on a 2D Concentric Circles Dataset. (a) Training dataset with 1,500 samples classified by their position relative to a unit circle (dashed line). Blue points represent the outer class, pink points the inner class. (b) Learned decision boundary and prediction probability heatmap from a 3-layer MLP (test accuracy: 0.983). The black contour shows the 0.5 decision boundary. (c-e) Invariant sets for three query points (black stars) with prediction values p. Orange points represent all input locations that yield identical predictions under the trained model, demonstrating the equivalence relation established by our framework. The invariant sets approximate level curves of the learned decision function, revealing the geometric structure of the model's decision space.

- **Limited optimization horizon**: CG constrains optimization to the number of diffusion steps, imposing additional restrictions on the process. Our empirical analysis demonstrates that this constraint is insufficient for achieving optimal results in invariant set generation.
- **Latent space complications**: Most state-of-the-art diffusion models Papers With Code [2024] are trained using the Latent Diffusion Model (LDM) approach. When conditioning generation on class labels rather than specific neuron activations, one must provide noisy intermediate encoded images $\mathcal{D}(\mathbf{x}_t)$ to the classifier. This requirement can be addressed by either training classifiers for each timestep t or using noisy estimates of $x_0(t)$, but the additional decoder step amplifies error propagation.

3.2.2. Infinite Optimization Approach

Given these limitations, we adopt an *Infinite Optimization* strategy, specifically adapting Algorithm 1 from Augustin et al. [2024]. This approach decouples the optimization process from the diffusion sampling steps, allowing for more flexible and thorough exploration of the invariant set while maintaining image quality and realism. The detailed algorithm specification is provided in Appendix .2.

3.3. Quality and Realism Assurance

Our approach ensures that generated images maintain high quality and realism through several mechanisms. We build upon industry-standard frameworks for synthetic image detection and guarantee that generated samples do not contain adversarial noise hidden in high-frequency patterns.

3.3.1. Frequency Domain Optimization

To address potential high-frequency artifacts, we perform frequency domain optimization that guides the generation process to encode meaningful signals in low-frequency bands—those visible to the human eye. Specifically, we introduce a low-pass filter $\mathcal{F}$ before the objective function $\mathcal{L}$ and measure deviation from the original measurement across different cutoff frequencies f_c.

This frequency-aware approach ensures that:

- Generated images appear natural to human observers
- Invariant set membership is achieved through semantically meaningful variations rather than imperceptible noise
- The generated samples maintain the visual characteristics expected from the underlying data distribution

The combination of infinite optimization with frequency domain constraints allows our method to generate diverse, high-quality samples from invariant sets while preserving both mathematical rigor and visual realism.

Chapter 4

Experiments

Chapter 5

Applications

Chapter 6

Discussion

Chapter 7

Conclusion

.1. Appendix

.2. Infinite Optimization Algorithm

This appendix provides the detailed algorithmic specification for our invariant set generation method, adapted from the infinite optimization approach. Unlike the original text-conditioned diffusion guidance, our algorithm is specifically designed for generating images that belong to the same invariant set as a given query point.

Algorithm 1 Invariant Set Generation via Infinite Optimization

```
Require: Loss function $\mathcal{L}$, Query point $\mathbf{x}^*$, Target value $\mathcal{L}(\mathbf{x}^*)$, Step budget $B$, Loss threshold $\tau$,
    Learning rate $\eta$, Step size $\lambda$, Low-pass filter $\mathcal{F}$
Ensure: Generated sample $x$ such that $\mathcal{L}(x) \approx \mathcal{L}(\mathbf{x}^*)$
$z_T \sim \mathcal{N}(0, I)$    ▷ Draw starting latent
$target\_value = \mathcal{L}(\mathbf{x}^*)$    ▷ Store target invariant value
for $t = 1, \dots, T$ do    ▷ Initialize time step-dependent variables
    $C_t = \emptyset$    ▷ No conditioning (unconditional generation)
end for
$optim$ = SGD($z_T$, lr $= \eta$) or Shampoo($z_T$, lr $= \eta$)    ▷ Define the optimizer
$step\_count = 0$    ▷ Initialize step counter
while $step\_count < B$ do    ▷ Optimization loop with budget
    $z = z_T$    ▷ Reset to starting latent
    for $t = T, \dots, 1$ do    ▷ Denoising loop
        with gradient_checkpointing():
            $z$ = LightningDiT_step($z, t$)    ▷ Diffusion update according to LightningDiT
    end for
    $x = \mathcal{D}(z)$    ▷ Decode final latent using VAE decoder
    $current\_value = \mathcal{L}(x)$    ▷ Calculate unfiltered objective value
    $x_{filtered} = \mathcal{F}(x)$    ▷ Apply low-pass filter
    $current\_value_{filtered} = \mathcal{L}(x_{filtered})$    ▷ Calculate filtered objective value
    $loss_1 = \|current\_value - target\_value\|^2$    ▷ Unfiltered invariant set loss
    $loss_2 = \|current\_value_{filtered} - target\_value\|^2$    ▷ Filtered invariant set loss
    $total\_loss = \lambda \cdot (loss_1 + loss_2)$    ▷ Combined loss with step size
    if $total\_loss < \tau$ then    ▷ Check convergence threshold
        break    ▷ Early termination
    end if
    $total\_loss$.backward()    ▷ Calculate gradients w.r.t. $z_T$
    $optim$.step()    ▷ Update starting latent
    $optim$.zero_grad()    ▷ Clear gradients
    $step\_count = step\_count + 1$    ▷ Increment step counter
end while
return $z_T, x$    ▷ Return optimized latent and final image
```

.2.1. Key Differences from Original Algorithm

Our adaptation introduces several important modifications to suit invariant set generation:

- **Unconditional Generation**: Unlike the original text-conditioned approach, we use unconditional diffusion models ($C_t = \emptyset$) and rely entirely on the optimization process to guide generation toward the target invariant set.
- **Invariant Set Objective**: Instead of optimizing for text-image alignment, we minimize the L_2 distance between $\mathcal{L}(x)$ and the target value $\mathcal{L}(\mathbf{x}^*)$, ensuring membership in the same invariant set.
- **Frequency Domain Filtering**: We incorporate a low-pass filter $\mathcal{F}$ before computing the objective function to ensure that invariant set membership is achieved through perceptually meaningful variations rather than high-frequency adversarial noise.

- **LightningDiT Integration**: The diffusion denoising process follows the LightningDiT sampling procedure, which may use different update rules than standard DDIM depending on the specific implementation and training configuration.

.2.2. Computational Considerations

The infinite optimization approach requires careful management of computational resources:

- **Gradient Checkpointing**: We employ gradient checkpointing during the denoising loop to reduce memory consumption while maintaining gradient flow through the entire diffusion process.
- **Optimizer Selection**: Based on empirical evaluation, SGD and Shampoo optimizers Gupta et al. [2018] demonstrate superior convergence properties for invariant set generation compared to adaptive methods like Adam.
- **Step Budget Management**: The algorithm balances computational cost with solution quality through the step budget B and threshold τ parameters, enabling early termination for efficient optimization landscapes.
- **Dual Loss Computation**: Computing both filtered and unfiltered objective values provides robustness against adversarial solutions while maintaining semantic coherence in generated samples.

Bibliography

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for investigating networks – uncovering classifier differences neuron visualisations and visual counterfactual explanations, 2024. URL https://arxiv.org/abs/2311.17833.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations, 2017. URL https://arxiv.org/abs/1704.05796.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Noa Cohen, Hila Manor, Yuval Bahat, and Tomer Michaeli. From posterior sampling to meaningful diversity in image restoration, 2024. URL https://arxiv.org/abs/2310.16047.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large ai models with semanticlens, 2025. URL https://arxiv.org/abs/2501.05398.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019. URL https://arxiv.org/abs/1902.03129.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization, 2018. URL https://arxiv.org/abs/1802.09568.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL https://arxiv.org/abs/1711.11279.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Deepdream – a code example for visualizing neural networks. https://research.google/blog/deepdream-a-code-example-for-visualizing-neural-networks/, 2015. Accessed: 2025-04-18.

Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere – large-scale detection of harmful spurious features in imagenet, 2023. URL https://arxiv.org/abs/2212.04871.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. doi: 10.23915/distill.00007. https://distill.pub/2017/feature-visualization.

Papers With Code. State-of-the-art: Image generation on imagenet 256x256. https://paperswithcode.com/sota/image-generation-on-imagenet-256x256, 2024. Accessed: April 6, 2025.

Oskar Pfungst. *Clever Hans (the Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*, volume 8. Holt, Rinehart and Winston, 1911.

Alin C. Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on Signal Processing*, 53(2):758–767, 2005. doi: 10.1109/TSP.2004.839932.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL https://arxiv.org/abs/1312.6034.

Bartlomiej Sobieski, Jakub Grzywaczewski, Bartlomiej Sadlej, Matthew Tivnan, and Przemyslaw Biecek. Rethinking visual counterfactual explanations through region constraint, 2024. URL https://arxiv.org/abs/2410.12591.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023. URL https://arxiv.org/abs/2303.09295.

Jingfeng Yao, Cheng Wang, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *Advances in Neural Information Processing Systems*, 37:56166–56189, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images, 2019. URL https://arxiv.org/abs/1907.06515.

Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023.