Titre: Title:	Building Robust Deep Learning Models for Visual Perception Tasks
Auteur: Author:	Seif Mzoughi
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Mzoughi, S. (2025). Building Robust Deep Learning Models for Visual Perception Tasks [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/63448/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/63448/
Directeurs de recherche: Advisors:	Foutse Khomh
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Building Robust Deep Learning Models for Visual Perception Tasks

SEIF MZOUGHI
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Mars 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Building Robust Deep Learning Models for Visual Perception Tasks

présenté par **Seif MZOUGHI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Giovanni BELTRAME, président
Foutse KHOMH, membre et directeur de recherche
Heng LI, membre

DEDICATION

To my family. . . ,

ACKNOWLEDGEMENTS

First, I would like to express my profound gratitude to Pr. Foutse Khomh, who provided me with the opportunity to enter the world of research and for his guidance throughout the Master's program, as well as for all of his feedbacks and support, which led to the realization of this thesis. Second, I want to thank to Dr. Mohamed Elshafei, whose feedback, knowledge, and research experience helped me progress in my research and enrich my knowledge. Thirdly, I'd like to express my gratitude to the lab mates with whom I've shared many moments. I would like especially to recognize Ahmed Haj Yahmed, with whom I have collaborated and who has been a great friend. I would like also to acknowledge Pr. Giovanni Beltrame and Pr. Heng Li for evaluating my thesis. Finally, I would like to thank my family for their support.

RÉSUMÉ

L'Apprentissage Profond (Deep Learning - DL) a révolutionné de nombreux domaines, transformant fondamentalement notre manière d'aborder les problèmes complexes en vision par ordinateur, traitement du langage naturel et systèmes autonomes. Toutefois, malgré des avancées spectaculaires, ces modèles restent vulnérables aux perturbations adversariales et aux entrées inattendues, ce qui remet en question leur fiabilité dans des applications critiques. Cette préoccupation croissante est particulièrement marquée en vision par ordinateur, où les défaillances des modèles peuvent avoir de graves conséquences dans le monde réel. Garantir leur robustesse implique de relever plusieurs défis fondamentaux : identifier précisément les vulnérabilités des modèles, diversifier les scénarios de test pour mieux évaluer leur résilience, s'assurer que les perturbations utilisées pour ces évaluations restent réalistes, et enfin, développer des stratégies efficaces pour renforcer la fiabilité des modèles face aux attaques et aux changements de distribution des données.

Dans cette étude, nous abordons ces défis sous deux angles complémentaires. D'abord, nous proposons une évaluation systématique des Générateurs d'Entrées de Test (Test Input Generators - TIGs) utilisés dans la classification d'images. Ces outils sont essentiels pour identifier les faiblesses des modèles en générant des entrées capables de révéler leurs limites. Cependant, leur efficacité varie selon plusieurs critères, notamment leur capacité à détecter des défauts, la diversité des entrées générées et l'authenticité des perturbations produites. Pour mieux comprendre ces aspects, nous analysons quatre TIGs de pointe, i.e., DeepHunter, DeepFault, AdvGAN et SinVAD; en les testant sur trois modèles pré-entraînés (LeNet-5, VGG16 et EfficientNetB3) et des ensembles de données de complexité croissante (MNIST, CIFAR-10 et ImageNet-1K). Nos résultats montrent que les approches basées sur des modèles génératifs, comme AdvGAN et SinVAD, sont particulièrement efficaces pour exposer des problèmes de robustesse sur des jeux de données simples, mais peinent à produire des résultats convaincants sur des distributions plus complexes. En revanche, les approches basées sur la perturbation, comme DeepHunter qui utilise le fuzzing guidé par la couverture et DeepFault qui exploite la localisation de fautes, offrent une meilleure stabilité et une performance plus homogène à travers différentes tâches. Ces observations soulignent la nécessité de concevoir des outils de test plus adaptatifs et capables de capturer la complexité croissante des ensembles de données du monde réel.

Deuxièmement, nous explorons la segmentation d'images, un domaine où la robustesse des modèles est encore peu étudiée malgré son importance dans des applications critiques telles

que l'imagerie médicale et la conduite autonome. Contrairement à la classification, où les fautes peuvent être détectées par des erreurs directes de prédiction, la segmentation exige des analyses plus fines, car les défaillances se traduisent par des erreurs localisées qui peuvent être difficiles à détecter avec des approches de test classiques. Pour pallier ce manque, nous introduisons le Test Métamorphique de Robustesse en Segmentation (SegRMT), une approche innovante combinant les tests métamorphiques avec des algorithmes génétiques afin de générer des entrées adversariales pertinentes. Testé sur le modèle DeepLabV3 avec le jeu de données Cityscapes, SegRMT démontre une capacité à détecter des vulnérabilités subtiles tout en améliorant la robustesse des modèles de manière plus efficace que les méthodes d'entraînement adversarial traditionnelles. Son innovation clé repose sur la génération de perturbations qui préservent la cohérence visuelle tout en exerçant un impact significatif sur les performances du modèle, grâce à une stricte maîtrise du rapport signal-bruit de crête (PSNR). Cette approche conduit à des améliorations notables en termes de généralisation, comme en témoignent les scores moyens plus élevés d'Intersection over Union (mIoU) lors de tests adversariaux croisés.

Notre travail apporte une contribution majeure à l'assurance qualité en apprentissage profond, en proposant à la fois un cadre d'évaluation exhaustif des TIGs existants et une technique novatrice pour renforcer la robustesse des modèles de segmentation. Ces outils et méthodologies permettent aux praticiens d'évaluer plus efficacement et d'améliorer la fiabilité des systèmes d'apprentissage profond dans des applications critiques, contribuant ainsi au développement de systèmes d'IA plus résilients et fiables.

ABSTRACT

Deep Learning (DL) has revolutionized numerous domains, fundamentally transforming how we approach complex problems in computer vision, natural language processing, and autonomous systems. However, despite remarkable progress, these models remain vulnerable to adversarial perturbations and unexpected inputs, calling into question their reliability in critical applications such as healthcare, autonomous driving, and cybersecurity. Ensuring their robustness involves tackling several fundamental challenges: precisely identifying model vulnerabilities, diversifying test scenarios to better assess their resilience, ensuring that perturbations used for evaluations remain realistic, and developing effective strategies to strengthen model reliability against attacks and shifts in data distribution.

In this study, we address these challenges from two complementary perspectives. First, we propose a systematic evaluation of Test Input Generators (TIGs) used in image classification. These tools are essential for identifying model weaknesses by generating inputs designed to expose their limitations. However, their effectiveness varies according to several criteria, including their ability to detect defects, the diversity of generated inputs, and the authenticity of the perturbations produced. To better understand these aspects, we evaluate four state-of-the-art TIGs, i.e., DeepHunter, DeepFault, AdvGAN, and SinVAD; testing them on three pre-trained models (LeNet-5, VGG16, and EfficientNetB3) and datasets of increasing complexity (MNIST, CIFAR-10, and ImageNet-1K). Our results show that generative model-based approaches, such as AdvGAN and SinVAD, are particularly effective at exposing robustness issues on simpler datasets but struggle to produce convincing results on more complex distributions. In contrast, more traditional approaches like DeepFault offer better stability and more consistent performance across different tasks. These findings highlight the need for more adaptive testing tools capable of capturing the growing complexity of real-world datasets.

Secondly, we explore image segmentation, an area where model robustness is still mostly underexplored despite its importance in critical applications such as medical imaging and autonomous driving. Unlike classification, where faults can be detected through direct prediction errors, segmentation requires more detailed analysis, as failures manifest as localized errors that can be challenging to detect with traditional testing approaches. To address this gap, we introduce Metamorphic Robustness Testing for Segmentation (SegRMT), an innovative approach combining metamorphic testing with genetic algorithms to generate relevant adversarial inputs. Tested on the DeepLabV3 model with the Cityscapes dataset, SegRMT

demonstrates the ability to detect subtle vulnerabilities while improving model robustness more effectively than traditional adversarial training methods. The key innovation of Seg-RMT lies in generating perturbations that maintain visual coherence while exerting a significant impact on model performance. This is achieved through a strict control of the Peak Signal-to-Noise Ratio (PSNR). This approach leads to significant improvements in terms of generalization, as evidenced by higher average Intersection over Union (mIoU) scores during cross-adversarial testing.

Our work makes an important contribution to quality assurance in deep learning by providing both a comprehensive evaluation framework for existing Test Input Generators (TIGs) and an innovative technique for enhancing the robustness of segmentation models. These tools and methodologies enable practitioners to more effectively assess and improve the reliability of deep learning systems in critical applications, thereby contributing to the development of more resilient and reliable AI systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

TIG	Test input generator
PSNR	Peak Signal to Noise Ratio
mIoU	mean intersection over usion
OOD	Out Of Distrbution Data
GMA	Generative Model Approaches
PBA	Perturbation-Based Approaches
GA	Genetic Algorithms
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning

CHAPTER 1 INTRODUCTION

Deep learning has fundamentally transformed modern technology, achieving remarkable success across diverse domains [1]. This success has driven a rapid adoption in safety-critical applications where reliability is paramount - from autonomous vehicles navigating complex urban environments [2] to medical imaging systems detecting life-threatening conditions [3]. As these systems become a vital part of our critical infrastructure, ensuring their robustness is not merely a technical challenge—it is now a crucial need. Nowadays, computer vision is one of the most demanding domains for adopting deep learning approaches in their application; hence, addressing their robustness concerns is crucial. The field includes diverse fundamental tasks with a special focus on classification and segmentation [4]. Classification models, which assign categorical labels to images, serve as the foundation for applications ranging from facial recognition [5] to industrial quality control [6]. Segmentation models perform the more complex task of pixel-level analysis, enabling applications like autonomous driving systems that must precisely identify road boundaries [2], or medical imaging tools that delineate tumor margins [7]. The growing real-world deployment of these systems has revealed distinct robustness challenges in each domain. While both classification and segmentation models must maintain reliability under varying real-world conditions, their failure modes and consequences differ significantly. A classification error might misidentify an object, but a segmentation error could misjudge critical boundaries in a medical scan or misidentify safe driving zones. These high stakes have driven the development of various testing approaches, yet significant gaps remain in our ability to systematically evaluate and ensure model robustness [8].In the following sections, we will explore the key challenges in Testing robustness in deep learning systems and examine the limitations of existing testing approaches.

1.1 Challenges in Testing Robustness in Deep Learning

The quest for robust computer vision systems faces several fundamental challenges that grow more complex as applications become more sophisticated. At their core, deep learning models exhibit unexpected sensitivity to input variations [9] - subtle changes that humans barely notice can trigger significant errors in model predictions. This sensitivity becomes particularly problematic in real-world deployments where environmental conditions constantly change. In classification tasks, a primary challenge lies in understanding how various types of perturbations affect model predictions [10]. As image complexity increases, the relationship between input changes and model behavior becomes increasingly opaque [11]. Testing

approaches must generate meaningful test cases that reveal potential weaknesses while remaining relevant to real-world scenarios. The proliferation of testing tools, each approaching the problem differently, has created an additional challenge: how to effectively evaluate and compare these diverse approaches. Segmentation tasks, on the other hand, face even more demanding challenges due to their pixel-level prediction requirements [12]. Beyond simple categorical accuracy, segmentation models must maintain precise spatial relationships across entire images [13]. In autonomous driving, for instance, even slight errors in boundary detection between road and obstacles could lead to catastrophic decisions [14]. Moreover, test cases must preserve semantic consistency - ensuring that generated examples maintain meaningful relationships between different image regions while still effectively testing model robustness. These fundamental challenges in deep learning robustness manifest differently across testing approaches, revealing specific limitations in current methodologies. As the complexity of visual tasks increases, traditional testing approaches increasingly struggle to balance effective robustness evaluation with maintaining meaningful test scenarios. Understanding these limitations is crucial for developing more effective testing strategies for both classification and segmentation tasks.

1.2 Limitations of Existing Test Input Generators (TIGs)

Current approaches to ensuring the robustness of computer vision systems face significant limitations that undermine their practical effectiveness. In the classification domain, the rapid proliferation of classification testing tools, while offering diverse methods, has led to an ecosystem where practitioners struggle to make informed decisions on which tools best suit their needs. Without standardized evaluation criteria across key dimensions such as: fault-revealing capability, test case diversity, and computational efficiency, tool selection often relies on incomplete information rather than robust empirical evidence [15].

These challenges in classification testing take on additional complexity when applied to segmentation tasks. Traditional gradient-based methods such as FGSM, PGD, and C&W, effective in classification, reveal significant limitations when applied to segmentation tasks [16]. Though these methods control perturbation magnitude through parameters like epsilon and alpha, they generate a limited range of test scenarios that may not accurately reflect the diverse challenges encountered in real-world segmentation applications.

Critical scenarios faced by systems deployed in real-world environments, such as fluctuating lighting conditions, sensor noise, and environmental variations, are often beyond the reach of these gradient-optimized perturbations. Furthermore, the segmentation domain currently lacks testing approaches that combine controlled perturbations with real-world variations.

While existing methods maintain image integrity via constrained optimization, they do not offer mechanisms for generating test cases that systematically explore how models handle the combined effects of adversarial perturbations and natural environmental variations. This gap is particularly concerning in safety-critical applications, such as autonomous driving or medical imaging, where models must remain robust against both intentional attacks and natural operational variations.

These limitations present tangible risks when deploying computer vision systems. In classification, the lack of systematic evaluation of Test Input Generators (TIGs) prevents practitioners from making evidence-based decisions on testing strategies. In segmentation, the narrow scope of current testing methods exposes models to unexpected failure modes in real-world deployments [17]. As these systems become increasingly embedded in critical applications, addressing these limitations through more comprehensive testing approaches becomes all the more urgent.

1.3 Research Statement

Our research addresses the challenges and limitations outlined above by developing a comprehensive evaluation framework that provides practitioners with a systematic approach to assess and compare testing strategies across multiple dimensions. In segmentation testing, we introduce SegRMT, a specialized framework that integrates metamorphic testing with genetic optimization. This innovative approach overcomes the limitations of existing gradient-based methods by generating test cases that better reflect real-world conditions while preserving semantic integrity.

Together, these research contributions advance testing methodologies to meet the increasing complexity and criticality of computer vision applications, particularly at a time when both technical demands and regulatory requirements necessitate urgent solutions. Notably, new regulatory frameworks, such as the EU AI Act, require rigorous pre-deployment testing [18].

Our work not only addresses current technical gaps but also provides organizations with the tools necessary to comply with emerging regulatory requirements for AI system validation.

1.4 Thesis Overview

This thesis presents two complementary contributions that address critical gaps in robustness testing:

1. **A Comprehensive Evaluation of Test Input Generators for Complex Visual**

Tasks. Building on the challenge of tool proliferation without standardization, we conducted a systematic evaluation of four state-of-the-art TIGs: DeepHunter, DeepFault, AdvGAN, and SinVAD. Our analysis uncovers a critical pattern: as visual tasks grow more complex, existing tools face a growing trade-off between testing effectiveness and maintaining visual coherence. This limitation becomes particularly pronounced with high-resolution images or intricate scenes, highlighting even greater challenges for segmentation testing.

2. **SegRMT: A Specialized Framework for Segmentation Testing**. Our classification study revealed that existing tools struggle to preserve complex visual relationships, a crucial requirement for segmentation tasks. While classification testing only requires retaining sufficient visual information for class identification, segmentation depends on the precise preservation of spatial relationships and semantic boundaries. This limitation, combined with our earlier findings on the shortcomings of current tools, motivated the development of SegRMT. This framework introduces two key innovations: a metamorphic testing methodology specifically designed to preserve pixel-level semantic relationships and a genetic algorithm that optimizes test case generation while maintaining visual coherence through carefully calibrated PSNR thresholds.

1.5 Thesis Contribution

This thesis makes the following four major contributions.

1. **Comprehensive Evaluation of Test Input Generators:** We conduct the first systematic evaluation of Test Input Generators (TIGs) across four critical dimensions: fault-revealing capability, naturalness, diversity, and efficiency. Our analysis provides practitioners with evidence-based insights for selecting appropriate testing tools based on their specific needs and constraints.

2. **Standardized Benchmarking Infrastructure:** We developed a comprehensive benchmarking platform that implements standardized metrics (ASR, DDR, LPIPS, Perturbation Magnitude) for evaluating TIGs. This infrastructure, available at [19], enables consistent and reproducible evaluation of testing tools, facilitating future research in robustness testing.

3. **SegRMT Framework:** We developed SegRMT, a segmentation testing approach that combines metamorphic testing with genetic optimization to achieve superior effectiveness than state of the art approaches. SegRMT only affects model performance by

6.4% mIoU while maintaining a PSNR of 24dB, significantly outperforming traditional gradient-based approaches that achieve only 21.7% to 8.5% mIoU.

4. **Robustness Enhancement:** We conducted extensive experiments to evaluate model robustness after adversarial training. Our results demonstrate that models trained with SegRMT-generated examples achieve 53.8% mIoU in cross-adversarial testing, outperforming traditional approaches that only manage 2-10% improvement. These results highlight SegRMT's effectiveness in enhancing model robustness against diverse adversarial perturbations.

The research work accomplished in this thesis led to the publication/submission of the following research papers:

- **Seif Mzoughi***, Ahmed Haj Yahmed*, Mohamed Elshafei, Foutse khomh, Diego Elias Costa, Towards Assessing Deep Learning Test Input Generators, *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE), 2025*
- **Seif Mzoughi***, Mohamed Elshafei, Foutse Khomh, Evaluating and Enhancing Segmentation Model Robustness with Metamorphic Testing in *Journal of Systems and Software (JSS), 2024*

1.6 Thesis Outline

The remainder of this thesis is organized as follows :

- Chapter 2: Provides an overview of the preliminary key concepts essential for comprehending the subsequent sections of the thesis. In this chapter, we will introduce the key concepts of deep learning, an introduction to the semantic segmentation tasks, and we also discuss the notion of robustness and adversarial vulnerability in addition to the common evaluation metrics.
- Chapter 3: Reviews the relevant literature.
- Chapter 4: Presents a comprehensive evaluation framework designed to assess and compare various TIGs across multiple dimensions offering valuable insights into their strengths and limitations.

- Chapter 5: Introduce SegRMT, a novel robustness testing framework for segmentation models that employs a metamorphic testing approach combined with genetic algorithms in order to efficiently test the segmentation models against realistic adversarial examples.
- Chapter 6: We present the conclusion of this thesis and discuss future work.

CHAPTER 2 BACKGROUND

Chapter Overview

This chapter introduces the foundational concepts and frameworks necessary to understand the challenges and solutions presented in this thesis. Section 2.1 presents the basic principles of deep learning and neural networks, detailing their architectures and training methodologies. Section 2.2 discusses advanced deep learning architectures, highlighting their applications in computer vision tasks and the contrasting robustness challenges faced by classification and segmentation models. Section 2.3 discusses model robustness and security, examining metrics such as PSNR and mIoU for evaluating the reliability of deep learning systems under perturbations. Section 2.3 reviews testing methodologies, including traditional white-box and black-box approaches, as well as advanced techniques like metamorphic testing and genetic algorithms, which address the limitations of conventional methods. Finally, Section 2.4 explores defense mechanisms, focusing on adversarial training and its unique challenges in ensuring robustness for segmentation models. Together, these sections provide a comprehensive understanding of the key concepts that underpin robustness testing in deep learning, setting the stage for the subsequent literature review.

2.1 Introduction to Deep Learning and Neural Networks

Deep learning has emerged as a transformative approach in artificial intelligence, enabling machines to learn from vast amounts of data and perform complex tasks. At the core of deep learning are artificial neural networks—computational models inspired by the structure and function of the human brain. These networks consist of layers of interconnected nodes, or neurons, that process data by applying linear transformations followed by non-linear activation functions.

2.1.1 Neural Network Fundamentals and Training

Artificial neural networks are composed of an input layer, one or more hidden layers, and an output layer. Each neuron in a layer receives inputs from the neurons of the preceding layer, computes a weighted sum of these inputs, adds a bias term, and then applies an activation function to produce an output. This output becomes the input for the next layer in the network.

Mathematically, the operation of a single neuron can be expressed as:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{2.1}$$

where:

- y represents the neuron's output,
- x_i are the input signals,
- w_i are the corresponding weights,
- b is the bias term,
- f is the activation function (such as the sigmoid function or ReLU).

The learning process involves adjusting the weights w_i and biases b to minimize a loss function L, which quantifies the discrepancy between the network's predictions and the actual targets. This optimization is typically achieved using algorithms like stochastic gradient descent (SGD) and backpropagation. The update rule for the parameters is given by:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta L(\theta^{(t)}) \tag{2.2}$$

where:

- θ represents the model parameters (weights and biases),
- η is the learning rate,
- $\nabla_\theta L(\theta^{(t)})$ denotes the gradient of the loss function with respect to the parameters at iteration t.

Through iterative training on large datasets, neural networks can learn complex mappings from inputs to outputs, capturing intricate patterns and relationships in the data. This capability allows them to perform tasks such as image recognition, natural language processing, and speech synthesis with remarkable accuracy.

2.1.2 Deep Learning Architectures

Deep learning architectures extend neural networks by adding more hidden layers, enabling the modeling of higher levels of abstraction. Some of the most influential architectures in deep learning include:

Convolutional Neural Networks (CNNs) [20]: Specifically designed for processing grid-like data such as images, CNNs employ convolutional layers that apply filters to local receptive fields. This structure captures spatial hierarchies of features, making CNNs highly effective for image-related tasks by detecting edges, textures, and more complex structures within images.

Recurrent Neural Networks (RNNs) [21]: Tailored for sequential data and time-series analysis, RNNs have connections that form directed cycles, allowing information to persist and enabling the network to exhibit temporal dynamic behavior. This makes RNNs suitable for tasks like language modeling and speech recognition.

Autoencoders [22]: Used for unsupervised learning of efficient codings, autoencoders learn to compress input data into a lower-dimensional representation and then reconstruct the output from this representation. This capability is useful for tasks such as dimensionality reduction and anomaly detection.

Generative Adversarial Networks (GANs) [23]: GANs consist of two networks—a generator and a discriminator—that are trained simultaneously. The generator creates synthetic data samples, while the discriminator evaluates them against real data. GANs have been used for image generation, style transfer, and data augmentation.

These architectures have enabled significant advancements in various domains, allowing models to achieve remarkable performance and often surpass human-level accuracy.

2.2 Model Robustness and Security

The deployment of deep learning models in critical applications has elevated robustness from a theoretical concern to a practical imperative [24]. Vision models must maintain reliable performance not only under ideal conditions but also when faced with real-world variations and potential adversarial inputs [25]. This requirement becomes particularly complex when considering the distinct ways robustness manifests in classification versus segmentation tasks. Classification models exhibit vulnerability patterns that stem from their architectural focus on global feature extraction [26]. When these models encounter perturbations, their behavior reveals interesting failure modes. A slight modification of texture patterns, imperceptible to humans, can cause a model to switch its prediction with high confidence [27]. For example, a self-driving car's classification system might misidentify a stop sign due to minor changes in lighting conditions or subtle modifications to its surface patterns [28]. These vulnerabilities often arise from the model's reliance on specific feature combinations that, while effective for standard inputs, can be disrupted through careful perturbation. Segmentation models face a

fundamentally different set of robustness challenges due to their pixel-level prediction requirements [29]. Unlike classification, where success is binary, segmentation models must maintain accuracy across every pixel while preserving complex spatial relationships [30]. Consider an autonomous driving scenario: a perturbation that causes minor boundary shifts in road edge detection could have catastrophic consequences, even if the overall scene interpretation remains largely correct. Segmentation vulnerabilities manifest in several interconnected ways: First, boundary regions show particular sensitivity to perturbations. Small changes near object boundaries can cascade into significant shifts in segmentation maps, affecting the precise delineation of different regions. Second, local perturbations can disrupt spatial consistency, creating fragmented or incoherent segmentations even when individual pixel classifications remain plausible [31]. Third, the holistic nature of segmentation means that changes in one image region can unexpectedly affect predictions in distant areas due to the model's global context understanding [32]. The assessment of model robustness requires metrics that capture both visual quality and prediction accuracy [33]. The Peak Signal-to-Noise Ratio (PSNR) serves as a fundamental metric for evaluating test case validity, quantifying how much a perturbed input deviates from the original input:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_i^2}{\text{MSE}} \right)$$

where MAX represents the maximum possible pixel value (typically 255 for 8-bit images), and MSE denotes the Mean Squared Error between the original and perturbed images:

$$\text{MSE} = \frac{1}{mn} \sum_i \sum_j \left[I(i,j) - K(i,j) \right]^2$$

where I and K are the original and perturbed images of size m×n respectively. For segmentation tasks, the mean Intersection over Union (mIoU) provides crucial insights into spatial prediction accuracy under perturbation. For each class c, the IoU is computed as:

$$\text{IoU}(c) = \frac{|A \cap B|}{|A \cup B|}$$

where A represents the predicted segmentation mask for class c and B represents the ground truth mask. The mIoU is then calculated as the average across all classes:

$$\text{mIoU} = \frac{1}{N} \sum_c \text{IoU}(c)$$

where N is the number of classes. The combination of these metrics enables comprehensive robustness assessment. PSNR ensures test case validity by quantifying input perturbation magnitude, while mIoU measures how well spatial relationships and class predictions are maintained under these perturbations. Values above 20 dB for PSNR typically indicate that test cases maintain semantic meaning [34], while mIoU directly quantifies the degradation in segmentation performance. Lower PSNR values risk introducing artificial artifacts that invalidate testing insights, while mIoU captures both boundary precision and classification correctness, proving particularly valuable for understanding how well segmentation models maintain spatial relationships when faced with perturbations. The unique challenges associated with evaluating and ensuring robustness have driven the development of specialized testing approaches, prompting our exploration of more advanced evaluation methodologies capable of effectively addressing vulnerabilities in both classification and segmentation tasks.

2.3 Testing Methodologies

The complexity of modern vision models demands sophisticated testing approaches to ensure their robustness. Traditional testing methodologies broadly divide into white-box and black-box approaches. Each offers distinct insights but faces unique limitations when applied to classification and segmentation tasks. White-box testing provides complete access to model internals, enabling detailed analysis of gradient flows, activation patterns, and feature representations [35]. For classification models, this approach can reveal which features most influence predictions and how perturbations propagate through the network. However, in segmentation models, white-box testing faces additional complexity due to the need to track spatial information preservation throughout the network. The computational overhead of analyzing modern architectures becomes prohibitive, particularly when examining spatial relationship maintenance across multiple layers. Moreover, insights gained from internal analysis often fail to translate directly to real-world performance scenarios. Black-box testing approaches the challenge of evaluating model robustness by treating models purely as input-output function [36]. Within the black-box testing domain, gradient-based techniques like Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attacks have emerged as standard approaches [37].These methods systematically generate adversarial inputs by following the gradient of the model's loss function:

$$x' = x + \epsilon \cdot \text{sign}\left(\nabla_x L(\theta, x, y)\right)$$

where x represents the original input, ϵ controls perturbation magnitude, and L denotes the loss function. While these methods prove effective for classification tasks, they show signif-

icant limitations in segmentation testing. Their focus on optimizing global loss values fails to capture the complex spatial consistency requirements crucial for segmentation robustness. A perturbation that successfully compromises a classification model might generate unrealistic or semantically inconsistent results when applied to segmentation tasks. Advanced testing strategies have emerged to address these limitations. Metamorphic testing introduces a framework for generating test cases without requiring explicit oracle values [38]. This approach defines relationships between inputs and outputs that should remain invariant under specific transformations. For vision tasks, these metamorphic relations might include:

$$\mathrm{MR}_1 : f(\mathrm{rotate}(x, \theta)) \approx \mathrm{rotate}(f(x), \theta)$$

$$\mathrm{MR}_2 : f(\mathrm{scale}(x, s)) \approx \mathrm{scale}(f(x), s)$$

where f represents the model function, and θ and s denote rotation angle and scale factor respectively. These relationships prove particularly valuable for segmentation testing, where maintaining semantic consistency under transformation is crucial. Genetic algorithms complement these approaches by providing systematic means of exploring the vast test input space [39]. Unlike gradient-based methods that follow fixed optimization paths, genetic algorithms discover diverse perturbation patterns through evolutionary search. The fitness function typically combines multiple objectives:

$$F(x') = \alpha \cdot E(x') + \beta \cdot Q(x')$$

where E measures prediction error, Q quantifies visual quality, and α, β are weighting factors. This multi-objective optimization allows simultaneous consideration of prediction impact, visual plausibility, and spatial consistency. The synthesis of these testing strategies reveals different aspects of model robustness, particularly important when evaluating segmentation models where traditional approaches may miss critical vulnerabilities. This comprehensive testing approach becomes essential for ensuring reliable performance in safety-critical applications, even though it does not fully resolve all inherent challenges.

2.4 Defense Mechanisms

Robustness is critical for deep learning models, especially segmentation models used in safety-critical applications such as medical imaging and autonomous driving. One widely adopted strategy is adversarial training, where models are exposed to adversarial examples during

training. In this approach, the loss function is modified as follows:

$$L = (1 - \alpha) \cdot L_{\text{standard}} + \alpha \cdot L_{\text{adversarial}}$$

Here, the parameter

$$\alpha$$

controls the balance between standard performance and enhanced robustness. For classification tasks, adversarial training primarily ensures correct categorical predictions under perturbations. In contrast, segmentation models introduce additional challenges due to the pixel-level granularity required. In segmentation, defenses must address:

- **Boundary Precision:** Even minor errors in object boundaries can significantly affect performance, for instance in precisely delineating structures in medical images.
- **Spatial Consistency:** The model must generate coherent predictions across the entire image rather than isolated categorical outputs.
- **Local and Global Feature Balance:** Segmentation models need to capture fine details (local features) while preserving overall scene context (global features).

These requirements create inherent trade-offs in training. Enhancing robustness may sometimes reduce boundary accuracy or compromise fine detail recognition, issues that often occur in applications like autonomous driving, where both road edge detection and the small objects recognition (e.g., traffic signs) are vital Furthermore, the generation of effective adversarial examples is essential for robust training. Traditional gradient-based perturbations often fail to simulate the full range of realistic variations encountered in segmentation tasks. This shortcoming has led to the development of more advanced techniques that specifically target:

- Preservation of spatial relationships,
- Maintenance of boundary consistency,
- Multi-scale feature robustness,
- Semantic coherence across regions.

These defense mechanisms, including adversarial training and its advanced variants, serve as techniques to enhance model robustness. They improve a model's ability to handle adversarial perturbations and natural input variations.

2.5 Chapter Summary

This chapter laid the groundwork for understanding deep learning robustness testing in computer vision systems. We started by exploring the fundamentals of neural networks and their architectural evolution, emphasizing how design choices influence model vulnerabilities in classification and segmentation tasks.

The discussion on model robustness introduced key metrics (PSNR and mIoU) that quantify both visual quality and prediction accuracy, establishing frameworks for evaluating robustness across different vision tasks. We then explored testing methodologies, progressing from traditional approaches to advanced strategies like metamorphic testing and genetic algorithms, demonstrating how conventional methods' limitations have driven the development of more sophisticated testing approaches. Finally, we examined defense mechanisms, particularly adversarial training, revealing the distinct challenges in building robust segmentation models compared to classification systems. This chapter establishes the context for our subsequent literature review, where we will analyze existing robustness testing approaches and their limitations.

CHAPTER 3 LITTERATURE REVIEW

Chapter Overview

This chapter provides a detailed review of the literature on robustness testing for deep learning systems, with a focus on the challenges, methodologies, and innovations in testing approaches for classification and segmentation tasks. Section 3.1 discusses the growing impact of deep learning across critical domains and the resulting robustness challenges, particularly in safety-critical applications. Section 3.2 highlights the specific robustness demands of computer vision tasks, contrasting classification and segmentation requirements. Section 3.3 explores the evolution of classification testing approaches, emphasizing the role of Test Input Generators (TIGs) such as perturbation-based and generative methods while addressing the limitations of current evaluation frameworks. Section 3.4 examines the unique challenges in testing segmentation models, including spatial consistency, boundary precision, and semantic coherence. Finally, Section 3.5 introduces complementary methodologies, including metamorphic testing and genetic algorithms, which have shown potential in enhancing robustness testing strategies for neural networks.

3.1 Deep Learning's Growing Impact and Robustness Challenges

The rapid adoption of deep learning across critical domains has fundamentally transformed modern technology while simultaneously raising serious concerns about system reliability. As Patel and Thakkar [40] highlight, deep learning has achieved remarkable success across diverse applications, from healthcare diagnostics to autonomous vehicles, often matching or exceeding human performance. This success has driven widespread deployment, including in safety-critical systems where reliability is paramount. However, this fast adoption has revealed significant vulnerabilities. A. Paleyes et al. [41] document numerous failures in deployed systems - from autonomous vehicles misidentifying critical obstacles to medical AI systems recommending incorrect treatments. These aren't merely technical glitches; they represent fundamental challenges in ensuring robust performance under real-world conditions. The authors' analysis of case studies reveals that many failures stem from models encountering scenarios that differ subtly from their training data, highlighting a critical gap between laboratory performance and real-world reliability. The stakes of these reliability issues have drawn regulatory attention. The work of Kelly et al. [18] describes how the EU AI Act now mandates rigorous pre-deployment testing for high-risk AI systems, including

specific requirements for robustness validation. This regulatory framework reflects growing recognition that traditional software testing approaches are insufficient for deep learning systems, whose behavior can be unpredictable under slight input variations. This push for reliability has sparked new approaches to safety assurance. Recent research ny Dalrymple et al. [42] proposes that achieving "guaranteed safe AI" requires moving beyond empirical testing to formal verification methods. Their work emphasizes that robust performance isn't just about accuracy metrics but demands provable guarantees about system behavior under various conditions. This represents a fundamental shift in how researchers think about AI system validation, particularly in domains where failures could have severe consequences. These insights highlight a clear trend: as deep learning systems become increasingly integrated into critical infrastructure, traditional reliability assurance methods are falling short. Ensuring robustness requires more advanced testing approaches, particularly in computer vision, where models must consistently perform across diverse and often unpredictable real-world conditions. This challenge is further amplified by the unique demands of different visual tasks, underscoring the need for a closer examination of computer vision's specific robustness requirements.

3.2 Computer Vision: A Domain Demanding Robust Solutions

Computer vision represents a critical frontier in the robustness challenge, with applications that directly impact human safety and well-being. As documented by Laad et al. [43], computer vision applications have expanded dramatically, now powering critical systems from medical diagnostics to autonomous navigation. This growth has revealed not just the technology's potential but also its vulnerabilities. Their analysis shows how these systems, while powerful, can fail in unexpected ways when confronted with real-world variations in lighting, perspective, or environmental conditions. The field naturally divides into two fundamental tasks - classification and segmentation - each presenting distinct robustness challenges. Bi et al. [44] trace the evolution of these tasks, showing how their different requirements lead to unique vulnerability patterns. Classification models, which assign categorical labels to entire images, can fail when subtle perturbations alter key features. In contrast, segmentation models must maintain pixel-level accuracy across entire images, making them vulnerable to both local and global disruptions. Chen et al. [45] reveal concerning patterns in classification failures, demonstrating how models that perform well on standard benchmarks can exhibit unexpected behaviors when faced with slight variations. Their analysis of class-wise vulnerabilities shows that certain categories consistently prove more challenging, suggesting systematic weaknesses in how models learn and generalize visual features. These findings

highlight the need for more nuanced approaches to testing and improving model robustness. The complexity of ensuring robustness becomes even more apparent in real-world deployments. Liu et al. [46] document how traditional robustness measures often fail to capture the full spectrum of challenges models face in practice. Their comprehensive study shows that current benchmarks and metrics can give an incomplete or misleading picture of model reliability, particularly when systems encounter scenarios that differ from their training distribution. This limitation becomes especially critical in segmentation tasks, where models must maintain both global coherence and local precision. Kamann et al. [47] report that segmentation models, despite high accuracy on standard metrics, can fail catastrophically when encountering real-world perturbations. Their analysis reveals that even state-of-the-art models show significant performance degradation under common image corruptions, highlighting the gap between benchmark performance and practical reliability. These challenges in computer vision robustness have led to the development of increasingly sophisticated testing approaches. However, as we'll explore in subsequent sections, the tools and methodologies developed for classification tasks often prove inadequate for segmentation, necessitating new approaches that can address the unique demands of pixel-wise predictions while maintaining semantic coherence.

3.3 Evolution of Classification Testing Approaches

The development of testing approaches for classification models reflects a growing understanding of deep learning vulnerabilities. Early testing methods focused primarily on accuracy metrics and basic robustness checks, but as shown by Ahuja et al. [48], these proved insufficient for ensuring reliable performance in real-world deployments. This limitation drove the development of more sophisticated testing tools, particularly Test Input Generators (TIGs) designed to systematically explore model behavior under various conditions. Two distinct approaches to test input generation emerged: perturbation-based and generative-based methods. Perturbation-based approaches, exemplified by DeepHunter and DeepFault, focus on systematic modification of existing inputs. DeepHunter, as described by Xie et al. [49], employs coverage-guided fuzzing with metamorphic mutations to generate test cases while preserving semantic meaning. This approach proved particularly effective at uncovering corner cases and boundary conditions that simpler testing methods missed. DeepFault, introduced by Eniser et al. [50], takes a different approach by using fault localization techniques to identify and target potentially problematic neurons, enabling more focused testing of model vulnerabilities. In parallel, generative approaches emerged to address limitations in perturbation-based testing. AdvGAN, developed by Xiao et al. [51], leverages generative

adversarial networks to create test inputs that challenge model robustness while maintaining visual realism. This approach marked a significant advance in test case generation, producing more natural perturbations than traditional methods. SINVAD, proposed by Kang et al. [52], further refined this approach by using variational autoencoders to navigate the space of valid inputs, enabling more systematic exploration of model behavior while preserving input validity. However, the proliferation of these tools created new challenges. As Maryam et al. [53] demonstrate, the lack of standardized evaluation frameworks made it difficult to compare different approaches effectively. Their analysis reveals significant inconsistencies in how TIGs are evaluated, making it challenging for practitioners to choose appropriate tools for their needs. This challenge is compounded by what Riccio and Tonella [54] identify as a critical gap in understanding when and why test generators produce invalid inputs, highlighting the need for more rigorous evaluation methods. The limitations of current evaluation approaches become more evident in real-world deployment. Different TIGs vary in effectiveness across models and datasets, yet without standardized comparison frameworks, these patterns remain poorly understood. This highlights the need for systematic evaluation methods to guide practitioners in tool selection and application.

More broadly, these challenges reflect a deeper issue in deep learning testing: the lack of comprehensive, standardized assessment frameworks. As we will see in the next section, this issue becomes even more complex in segmentation tasks, where pixel-wise predictions add new dimensions to the problem.

3.4 Unique Challenges in Segmentation Testing

While classification testing tools have matured significantly, their application to segmentation reveals a fundamental mismatch in testing requirements. As chen et al. [45] demonstrate, these tools operate on an inherently simpler premise - testing categorical predictions - while segmentation demands evaluation of complex spatial relationships across thousands of coordinated pixel-level decisions. This isn't merely a matter of scale; it represents a qualitatively different testing challenge that questions the very foundations of existing testing approaches. The stakes of this testing gap become clear in critical applications. In medical imaging, a misplaced boundary in tumor segmentation can mean the difference between successful treatment and serious complications. In autonomous driving, even minor errors in segmenting pedestrian boundaries can lead to catastrophic safety failures. Xu et al. [55] documents how these applications are rapidly expanding, making the testing gap increasingly urgent. Each domain, including medical image analysis, autonomous systems, and industrial robotics, introduces unique requirements for robustness testing - from maintaining precise boundary

detection under varying imaging conditions in medical scenarios to ensuring reliable performance across diverse environmental conditions in autonomous systems. Classification testing tools struggle with segmentation due to their underlying assumptions. They treat predictions as independent events, optimizing perturbations to flip categorical labels. In contrast, segmentation predictions are interdependent—an improvement in one region may compromise semantic consistency elsewhere. This interdependence undermines the mathematical foundations that make classification testing effective. Some and Namboodiri [56] identify several critical failure modes specific to segmentation models that illustrate this interdependence:

- **Boundary Degradation**: Models often show increased vulnerability at object boundaries, where slight perturbations can cause significant shifts in segmentation maps, creating ripple effects across the entire prediction.
- **Spatial Inconsistency**: Local perturbations can trigger non-local effects, causing segmentation errors in distant image regions due to the model's holistic understanding of scene context.
- **Class Confusion**: Similar visual patterns can lead to systematic misclassification, particularly for smaller or less frequent objects, with errors propagating across semantically related regions.
- **Scale Sensitivity**: Models exhibit varying robustness across different object scales, with smaller objects particularly vulnerable to perturbations that can completely eliminate their detection.

Current testing approaches show significant limitations when addressing these challenges. Arnab et al. [57] show that traditional adversarial testing methods, though effective for classification, fail to capture complex spatial dependencies in segmentation. Their study highlights a critical issue: segmentation models can appear robust on standard metrics while exhibiting vulnerabilities in spatial consistency and boundary precision—weaknesses that only emerge under specific perturbations. Recently, Halmosi et al. [58] showed that even models trained with standard adversarial techniques remain vulnerable to carefully crafted perturbations that exploit these segmentation-specific weaknesses. These perturbations often preserve global image statistics while disrupting crucial local spatial relationships, revealing blind spots in current testing approaches. Attempts to adapt classification testing methods to segmentation have had limited success. Gu et al. [59] show that while traditional approaches like PGD can be modified for segmentation, they face a key limitation: their

optimization objectives fail to capture the balance between local accuracy and global semantic consistency. Even when these methods generate adversarial examples, the perturbations often lack semantic meaning in a segmentation context, resulting in test cases that do not reflect realistic failure modes. This systematic analysis reveals a critical gap in the field. While classification testing has benefited from years of tool development and refinement, segmentation testing lacks specialized frameworks that can address its unique challenges. This thesis aims to fill this gap by proposing SegRMT, a specialized testing framework that integrates metamorphic testing with GA in order to generate adversarial examples designed to test the robustness of segmentation models. By systematically exploring the adversarial search space and balancing the local perturbations with global sematic integrity, SegRMT is designed to expose realistic failure modes in segmentation models. This approach not only fills the current testing gap by directly addressing the unique challenges of segmentation but also paves the way for improved model reliability in critical applications like medical imaging and autonomous driving.

3.5 Metamorphic Testing and Genetic Algorithms

The development of robust and reliable software systems has necessitated innovative testing methodologies. Among these, metamorphic testing and genetic algorithms have emerged as impactful techniques in software engineering and optimization, respectively. This section examines metamorphic testing and its foundational role in addressing the oracle problem in traditional software testing, its applications in deep learning, and the use of genetic algorithms in optimization and adversarial example generation for neural networks.

3.5.1 Metamorphic Testing in Software Engineering

Metamorphic testing is a software testing approach introduced by Chen et al [38] to alleviate the oracle problem—a situation where determining the correct output of a program for a given input is difficult or impossible. Instead of relying on test oracles to verify program outputs, metamorphic testing focuses on identifying relationships between inputs and outputs, known as metamorphic relations (MRs). These relations describe expected transformations in outputs when inputs are modified, providing a means to detect inconsistencies that indicate potential defects.

For instance, in a numerical computation program, a MR might state that scaling the input by a factor should proportionally scale the output. If this relationship does not hold, it suggests a fault in the program. Metamorphic testing has proven effective across domains

such as scientific computing, simulation software, and machine learning algorithms (Segura et al. [60]). It is particularly valuable for detecting subtle bugs in systems where the correctness of individual outputs is hard to ascertain, but the relationships between inputs and outputs are well understood.

Despite its strengths, metamorphic testing faces challenges, particularly in identifying and constructing effective MRs, which often require domain-specific knowledge. This process can be labor-intensive and inconsistent due to the lack of formal descriptions for many MRs (Segura et al. [60]). Nonetheless, its ability to uncover faults in complex systems where traditional methods fall short has made it very important in software engineering.

3.5.2 Applications in Deep Learning

Metamorphic testing has been applied to deep learning, where MRs define expected model behaviors under specific transformations. For instance, in image classification, an MR might state that a slight rotation of an image should not alter its predicted class. By verifying model adherence to such relations, metamorphic testing helps uncover subtle defects and vulnerabilities. Tian et al [61] applied metamorphic testing to convolutional neural networks and identified misclassifications caused by subtle input variations that escaped detection through traditional testing methods. However, scaling metamorphic testing for deep learning remains challenging, especially with high-dimensional data and complex architectures. Manually defining and implementing MRs is labor-intensive, and the lack of standardized tools for deep learning further hinders its adoption (Chen et al. [62]).

Another critical issue is balancing the specificity and generality of MRs. If an MR is overly general, it may fail to detect faults, whereas highly specific MRs might not apply to a wide range of inputs or could miss other types of faults. These challenges highlight the need for research into automated MR generation and tool support, which could enhance the applicability and scalability of metamorphic testing in deep learning. Despite these limitations, metamorphic testing remains a promising method for improving the robustness of neural networks, particularly when integrated with other testing techniques.

3.5.3 Genetic Algorithms in Robustness testing

Genetic algorithms (GAs) are adaptive heuristic search algorithms inspired by the process of natural selection and genetics. Introduced by Holland [63], GAs are widely used to solve optimization and search problems by iteratively evolving a population of candidate solutions towards better solutions based on a fitness function.

In a GA, a population of candidate solutions (individuals) is initialized, and genetic operators such as selection, crossover (recombination), and mutation are applied to evolve the population over successive generations. The fitness function evaluates how close a given solution is to the optimum. Individuals with higher fitness have a higher probability of being selected for reproduction, allowing advantageous traits to propagate through the population.

GAs are especially effective for optimization problems with large, complex, or poorly understood search spaces. Their robustness against local optima allows them to explore the solution space efficiently and identify near-optimal solutions.

GAs have been used to generate adversarial examples for neural networks, especially in black-box attack scenarios where the attacker has little to no knowledge of the model's architecture or parameters. Alzantot et al. [64] introduced a GA-based approach to generate adversarial examples for text classification models without requiring access to gradients.

An improved genetic algorithm (IGA) was introduced by D. Yang et al. [65] to enhance the efficiency and precision of traditional GAs in generating adversarial examples. The IGA incorporated modifications to the crossover and mutation operations. Specifically, it evaluated all possible crossover points and selected the best result to accelerate convergence. Mutation rates were dynamically adjusted to improve diversity and global search capabilities without sacrificing stability.

When applied to adversarial attack generation, the IGA effectively produced high-confidence adversarial examples against a deep convolutional neural network trained on the MNIST dataset. Operating in a black-box setting, it required only the target model's output classifications and confidence scores. Compared to traditional optimization methods such as particle swarm optimization and the grey wolf optimizer, the IGA achieved faster convergence and higher success rates (D. Yang et al. [65]). The use of genetic algorithms in adversarial example generation highlights their versatility and effectiveness in optimization problems involving neural networks. However, challenges persist, including the high computational cost of evaluating fitness functions, especially in high-dimensional search spaces. Moreover, GAs can be slower per iteration than gradient-based methods, as they require assessing multiple candidate solutions simultaneously.

3.6 Chapter Summary

This chapter examines the evolution of robustness testing in deep learning, focusing on the distinct challenges presented by classification and segmentation tasks. While classification models prioritize categorical accuracy, segmentation models must also preserve spatial rela-

tionships and semantic coherence, which significantly influences testing methodologies.

The chapter reviews the progression of classification testing, from basic perturbation techniques to more sophisticated Test Input Generators (TIGs) such as DeepHunter, DeepFault, AdvGAN, and SINVAD. Despite these advancements, the absence of standardized evaluation frameworks complicates the selection of appropriate tools. Although metamorphic testing and genetic algorithms provide useful approaches, they struggle to maintain spatial coherence in segmentation tasks.

Our literature review highlights a critical gap in current methodologies, as they fail to address the unique demands of segmentation, particularly in maintaining boundary precision and spatial consistency, which are vital in safety-critical applications such as medical imaging and autonomous driving. This chapter identifies two significant gaps: first, the need for systematic evaluation of TIGs for classification tasks, and second, the requirement for specialized testing approaches tailored to segmentation. These gaps motivate the development of both a comprehensive TIG evaluation framework and the SegRMT approach for segmentation testing.

CHAPTER 4 A COMPREHENSIVE EVALUATION OF TEST INPUT GENERATORS FOR COMPLEX VISUAL TASKS

Chapter Overview

This chapter presents a comprehensive empirical evaluation of four state-of-the-art TIGs (i.e., DeepHunter, DeepFault, AdvGAN, and SinVAD) across four key dimensions: fault-revealing capability, naturalness, diversity, and efficiency. We highlight trade-offs between the TIGs in terms of robustness detection, test case variation, and efficiency, revealing significant performance shifts based on dataset complexity, with each tool exhibiting varying behavior across different evaluation scenarios.

4.1 Context of the Study

Deep Learning (DL) has emerged as a transformative solution in computer vision, extracting complex patterns from diverse data sources [66–68]. DL components are now integral to safety-critical systems such as self-driving cars [69] and aircraft collision avoidance systems [70], where reliability and robustness are paramount.

However, despite their growing adoption, DL systems struggle to maintain consistent performance under diverse conditions. A key concern is robustness. For instance, an autopilot system's fatal accident [71] occurred due to failure in handling lighting conditions different from its training data. Other DL system failures stem from different issues: a facial recognition system's misidentification [72] highlighted accuracy limitations, while hiring system bias against women [73] revealed fairness concerns. These incidents collectively demonstrate that high accuracy on standard datasets does not guarantee reliable real-world performance, with robustness being a particularly critical concern for safety-critical applications. To address these robustness challenges, recent efforts have focused on developing test input generators (TIGs) [74–76]. TIGs are specialized techniques designed to generate new test inputs—either by modifying existing data or creating novel instances—that rigorously evaluate and stress-test DL models. They expose vulnerabilities unique to DL systems, such as sensitivity to minor input perturbations and susceptibility to adversarial examples, which traditional test case generation methods fail to uncover. TIGs specifically target these weaknesses by crafting inputs that can cause the model to produce erroneous outputs, thereby revealing robustness issues.

Over the years, numerous TIGs have been developed, leveraging a wide range of strategies

including (1) pixel-level perturbation [77]; (2) manipulation of the input representation using generative DL models such as generative adversarial networks (GANs) [78] or variational autoencoder (VAE) [79]. These TIGs vary widely in their methodologies and focus areas. Some focus on applying subtle perturbations to existing data, others generate entirely new data instances, and each introduces unique techniques to challenge and evaluate the robustness of DL models in different ways.

Despite the prevalence of TIGs, several challenges persist in the field. Researchers and developers now have access to a broad spectrum of tools, each operating differently, yet there has been no comprehensive assessment of TIGs across various models and datasets. This lack of evaluation hinders the generalization of their effectiveness and makes it challenging for practitioners to select the most suitable tool for their specific needs. The community requires deeper insights into these TIGs to better understand their capabilities and limitations.

Furthermore, clear assessment criteria for evaluating TIGs remain undefined. Many studies focus on generating failure-inducing cases without adequately considering the quality of the generated data. Assessing this quality, especially in terms of naturalness [80], is a significant challenge. Simple rule-based or distance-based metrics often fail to capture the nuances of human perception, leading to test cases that may not reflect real-world scenarios and thus limiting their practical utility. Additionally, efficiency in test case generation is often neglected, even though it is crucial for real-world applicability. Many TIGs are computationally intensive and resource-demanding, which hampers their use in practical settings. These challenges highlight the need for more comprehensive evaluation criteria that consider not only the fault-revealing capabilities of TIGs but also the naturalness, diversity, and efficiency of the generated test cases.

In this Chapter, we undertake the first comprehensive assessment of TIGs for DL systems, aiming to identify and understand their effectiveness across multiple dimensions: fault-revealing capability, naturalness, diversity, and efficiency. This chapter guides selecting appropriate TIGs based on specific testing needs and resource constraints while establishing a framework for assessing and improving testing approaches. To guide our investigation, we formulate our research questions as follows:

- **RQ1:** Which TIG **reveals** more DL robustness issues?
- **RQ2**: Which TIG generates more **natural test cases**?
- **RQ3**: Which TIG generates more **diversified test cases**?
- **RQ4:** Which TIG is more **efficient** in test case generation?

We conduct an empirical study evaluating four state-of-the-art TIGs: DeepHunter [76], DeepFault [50], AdvGAN [81], and SinVAD [82]. To ensure the generalizability of our findings, we leverage three DL models of varying complexities (LeNet-5 [83], VGG16 [84], and EfficientNetB3 [77]) applied to three datasets of different sizes (MNIST [83], CIFAR-10 [85], and ImageNet-1K [86]). We assess the TIGs on four dimensions: fault-revealing capability, naturalness [80], diversity, and efficiency. These dimensions represent the critical aspects that determine a TIG's practical value [87]. In fact, fault-revealing capability demonstrates effectiveness in identifying real issues [75], naturalness ensures that test cases reflect realistic scenarios [80], diversity guarantees comprehensive testing coverage [88], and efficiency enables practical deployment. Based on this evaluation framework, we aim to provide practical insights for both the software engineering (SE) and artificial intelligence (AI) communities on the selection and application of TIGs for robustness testing.

The remainder of this chapter is organised as follows: Section 4.2 describes the study design. Section 4.3 summarizes the evaluation findings across research questions RQ1 to RQ4. Section 4.4 presents a discussion of the results, Section 4.5 examines threats to the study's validity. Finally, Section 4.6 concludes the chapter.

4.2 Study Design

In this study, we evaluated DL TIGs using a three-step methodology: (1) selection of representative pre-trained models, datasets, and TIGs; (2) use of TIGs to generate synthetic test inputs; and (3) evaluation of the TIGs. Figure 4.1 illustrates this workflow. First, we selected four well-known TIGs that represent the main categories in the literature: DeepHunter [89], DeepFault [50], SinVad [82], and AdvGAN [81]. For DL models, following previous comprehensive evaluation studies of DL testing approaches that assess multiple TIGs across varied model architectures and datasets [15], we chose three popular pre-trained models with varying levels of complexity: LeNet-5 [83] (small complexity), VGG-16 [90] (medium complexity), and EfficientNetB3 [91] (large complexity). To evaluate these models using TIGs, we employed three widely used publicly available datasets of different sizes: MNIST [92] (small size), CIFAR-10 [85] (medium size), and ImageNet1K [93] (large size).

We then applied each TIG to each pre-trained DL model to generate new input samples to test model robustness. Finally, we assessed each TIG using four key metrics: Defect Detection Rate (DDR) [76], Attack Success Rate (ASR), Learned Perceptual Image Patch Similarity (LPIPS) [94], and perturbation magnitude (PM) [95]. Each experiment was repeated ten times, and all values were recorded. We used averages for summarizing results and the raw values to conduct statistical significance tests on the differences between means.

The following sections describe the datasets selected, TIGs used for comparison, the DL models employed, and the metrics used in our experiments.

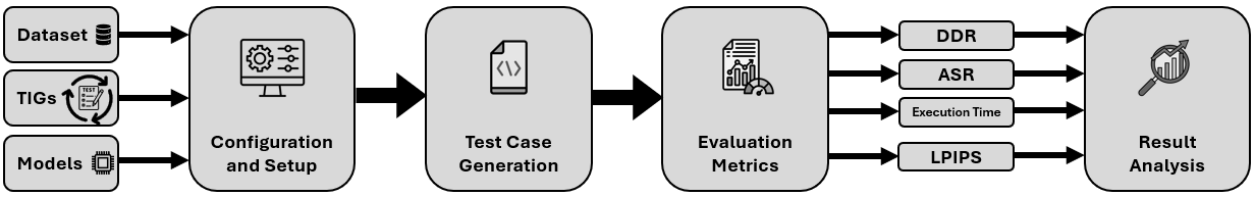


Figure 4.1 Study Methodology for Assessing Test Input Generator (TIG) Performance

4.2.1 Models and Datasets

We selected three widely recognized model-dataset pairs to assess TIGs, representing different complexity scales in DL. We chose LeNet-5 and VGG16, foundational architectures designed for image classification tasks. LeNet-5 processes 28×28 grayscale images and is paired with the MNIST dataset, which contains 60,000 training and 10,000 test images of handwritten digits across 10 classes. VGG16, a deeper Convolutional Neural Network (CNN) with 16 layers, handles $32 \times 32 \times 3$ RGB images and is suited for the CIFAR-10 dataset, comprising 50,000 training and 10,000 test images across 10 classes. To incorporate models with higher complexity, we included EfficientNetB3 from the family of efficient architectures [77] using compound scaling [96]. EfficientNetB3 processes bigger resolution images of size $224 \times 224 \times 3$ and is paired with the ImageNet1K dataset. ImageNet1K offers a collection of 1.2 million training images and 50,000 validation images across 1,000 classes, making it a standard benchmark for image classification tasks. These dataset-model pairs represent evaluation subjects with increasing complexity and resolution: MNIST (28×28 grayscale), CIFAR-10 (32×32 RGB), and ImageNet-1k (300×300 RGB), as detailed in Table 4.1. This progression allows us to evaluate the robustness of the TIGs across diverse data distributions and feature richness.

These combinations were selected based on three criteria: (1) established performance benchmarks in the literature, (2) varying architectural complexity to test TIG scalability, and (3) diverse image characteristics to evaluate TIG adaptability. All models used are pre-trained versions that achieve competitive accuracy on their respective datasets.

4.2.2 DL Test Input Generators

We selected the four TIGs outlined in Section 3 because they encompass (1) diverse test generation approaches, i.e., PBA and GMA; (2) different access levels, i.e., white-box and

Table 4.1 Comparison of Test Input Generators (TIGs)

Dataset	Model	Resolution	Training Set	Test Set	Classes	Sample Size
MNIST	LeNet-5	28×28 (Gray)	60,000	10,000	10	1,000 (10%)
CIFAR-10	VGG16	32×32 (RGB)	50,000	10,000	10	1,000 (10%)
ImageNet-1K	EfficientNetB	300×300 (RGB)	1,200,000	50,000	1,000	1,000 (2%)

black-box; (3) various test objectives, i.e., including misbehaviors, neuron coverage, and surprise coverage; and (4) they are open-source, allowing us to execute and evaluate them.

We applied modifications to the selected TIG when necessary to ensure a fair and comprehensive comparison. For the ImageNet1K [97] experiments, three TIGs required adjustments: For DeepHunter, we modified the profiler to handle higher-resolution images (300 × 300 pixels) and EfficientNet-specific layers. We also updated the preprocessing steps to match EfficientNet's input format and adjusted the layer selection to skip layers with less informative neuron coverage, allowing DeepHunter to effectively capture relevant activations in this model. DeepFault was adapted to handle higher-resolution images (300 × 300 pixels) and to interface with EfficientNetB3, while AdvGAN's generator network was scaled to handle the higher-resolution input of EfficientNetB3, ensuring compatibility without changing its core structure. Despite similar efforts to adapt SINVAD's Variational Autoencoder (VAE) architecture to accommodate the higher resolution required by EfficientNetB3, it proved insufficient to handle the complexity of ImageNet1K. SINVAD's VAE architecture could not be successfully adapted to process the higher-resolution ImageNet1K inputs. For the experiments involving MNIST and CIFAR-10 datasets, all tools retained their original implementations.

4.2.3 Generation of Synthetic Test Inputs

In this step, for each of the pre-trained DL models and the datasets described in section 4.2.1, we leverage the TIGs outlined in Subsection 4.2.2 to generate synthetic test inputs. From each dataset's test set, we create 10 non-overlapping folds of 100 samples each, ensuring that all selected samples are correctly classified by the model under test. Each fold is treated as a separate experiment, maintaining independence between folds. For each fold, we apply the TIGs to generate synthetic test cases. For each TIG, we construct a new synthetic collection of test cases, D_{syn}, where the number of generated images varies according to each TIG's

generation strategy. D_{syn} is then used to evaluate the performance of the TIGs. Some TIGs, such as DeepHunter and DeepFault, can be executed with different configurations, i.e., DeepHunter can operate using six different coverage criteria, while DeepFault can utilize three different suspiciousness measures (Ochiai, DStar and Tarantula). We run these TIGs across all configurations and record the average values. This experimental design aligns with the guidelines in “Empirical Standards for Software Engineering Research" [98]. In practice, when comparing approaches that incorporate random components, it is essential to perform multiple runs and apply statistical significance tests. The inherent randomness of these TIGs necessitates several runs to validate the tests' outcomes. For parameter tuning in each approach, we maintain the default values provided by the respective TIGs. We acknowledge that exploring the impact of different parameter settings in each approach represents an interesting direction for future work.

4.2.4 Evaluation Criteria

This section describes the metrics used to evaluate the generated inputs and assess the effectiveness of the TIGs. We employ five complementary metrics: Defect Detection Rate (DDR) [50] [89], Attack Success Rate (ASR) [99], Learned Perceptual Image Patch Similarity (LPIPS) [94], Perturbation Magnitude (PM) [100], and Execution Time (ET).

Defect Detection Rate (DDR). The DDR quantifies a TIG's ability to expose robustness issues by calculating the proportion of original test inputs that, after transformation, lead to misclassification. DDR measures the TIG's capability to make original inputs challenging for the model, providing insight into the overall robustness of the model when exposed to transformed examples. Formally, the DDR is defined as:

$$\text{DDR} = \frac{N_{\text{misclassified original}}}{N_{\text{total original}}} \tag{4.1}$$

where $N_{\text{misclassified original}}$ is the number of original test inputs that result in misclassification after transformation by the TIG, and $N_{\text{total original}}$ is the total number of original test inputs.

Attack Success Rate (ASR). The ASR measures the success rate of the adversarial examples generated by the TIGs in causing misclassification. ASR focuses on the effectiveness of the adversarial examples produced, providing insight into how often the generated examples successfully cause misclassification. Formally, the ASR is defined as:

$$\text{ASR} = \frac{N_{\text{successful adversarial}}}{N_{\text{total generated}}} \tag{4.2}$$

where $N_{\text{successful adversarial}}$ is the number of adversarial examples that result in misclassification, and $N_{\text{total generated}}$ is the total number of generated test inputs. While both metrics assess a TIG's ability to induce misclassification, they focus on different aspects. DDR measures the proportion of original test inputs for which the TIG can find at least one misclassification. It emphasizes coverage across all original inputs, evaluating how effectively the TIG challenges each one. ASR, in contrast, calculates the overall success rate of all generated test inputs in causing misclassification, regardless of how many original inputs they originate from. By comparing DDR and ASR, we understand both the TIG's thoroughness in covering the test set (DDR) and the effectiveness of its generated inputs in inducing misclassification (ASR).

Learned Perceptual Image Patch Similarity (LPIPS). To evaluate the naturalness and quality of the generated test cases, we employ LPIPS [94] metric. LPIPS calculates the perceptual distance between images based on deep features extracted from a neural network, providing a semantically meaningful assessment of image similarity. Specifically, we utilize the AlexNet architecture [97] for feature extraction due to its effectiveness in capturing perceptual differences. This metric ensures that the generated test cases are perceptually distinct from the original inputs while avoiding nonsensical images that are not useful for robustness testing.

Perturbation Magnitude (PM). The PM measures the extent of changes applied to the original inputs to create synthetic inputs. It is calculated as the mean L2 norm of the difference between the original and transformed images, offering a quantitative assessment of the distortion introduced by the TIGs. Formally, PM is defined as:

$$\text{PM} = \frac{1}{N} \sum_{i=1}^{N} \left\| x_i^{\text{orig}} - x_i^{\text{trans}} \right\|_2 \tag{4.3}$$

where N is the total number of pixels in the image, x_i^{orig} is the value of the ith pixel in the original image, and x_i^{trans} is the corresponding pixel value in the transformed image.
In our experiments, we compute (1) the average PM (Avg PM) and (2) the standard deviation PM (Std PM) across all images. The Avg PM indicates the overall level of change introduced by the TIG, while the Std PM reflects the diversity of the TIG transformations.

Execution Time (ET). To further evaluate the TIGs, we measured the execution time per image, calculated as the total processing time divided by the number of images tested. This metric captures the efficiency of each TIG, allowing us to assess their practicality in real-time or resource-limited scenarios.

By combining DDR, ASR, LPIPS, PM, and ET, we provide a comprehensive evaluation of the TIGs. The DDR and ASR assess the TIGs' ability to reveal robustness issues and

generate effective adversarial examples, respectively. The LPIPS assesses the perceptual similarity between the original and generated inputs, evaluating the quality of the generated test inputs. The PM quantifies the level of perturbation introduced to the original inputs, providing a quantitative assessment of the extent and diversity of transformations applied. ET captures the efficiency of each TIG.

4.3 Results

In this section, we first briefly introduce the experimental environment, and then we detail the experiments and results obtained to answer our RQs. All experiments were conducted on Compute Canada's infrastructure, leveraging Intel E5-2683 v4 Broadwell processors and NVIDIA A100 GPUs running on a Linux operating system.

4.3.1 RQ1: Which TIG reveals more DNN robustness issues?

Motivation. The goal is to evaluate the effectiveness of state-of-the-art TIGs in revealing DNN robustness issues by identifying misclassified synthetic test cases, thus exposing potential vulnerabilities in the DL models.

Method. We generated synthetic test cases for each TIG from the original test set and computed the average Defect Detection Rate (DDR) and Attack Success Rate (ASR) across 10 runs. Since we are working with averaged metrics, to assess whether the differences across approaches are statistically significant, we use the Wilcoxon statistical test [101] and Vargha-Delaney [102] ($\hat{A}_{12}$) effect size test. The Wilcoxon statistical test determines if the difference between two means is statistically significant (p-value < 0.05). Vargha-Delaney $\hat{A}_{12}$ determines the magnitude of difference between two groups, with a range of [0, 1]. $\hat{A}_{12} > 0.5$ implies that values in the first group are larger, $\hat{A}_{12} < 0.5$ implies that they are lower, and $\hat{A}_{12} = 0.5$ indicates statistically indistinguishable groups. We used the coefficients proposed by Hess et al. [103] to interpret the magnitude of the differences into negligible, small, medium, and large differences.

Results: Table 4.2 summarizes DDR and ASR results across datasets. Results reveal clear differences in the robustness-revealing capabilities of TIGs across datasets and models.

For MNIST, GMA TIGs (AdvGAN and SinVAD) significantly outperformed PBA TIGs (DeepHunter and DeepFault). AdvGAN achieved the highest DDR/ASR (99.1%), followed by SinVAD (89.0%), due to their generative capabilities. In contrast, PBAs like DeepHunter had moderate performance (DDR: 71.4%, ASR: 38.24%), limited by their small-scale perturbations. DeepFault exhibited the weakest performance (DDR/ASR: 1.0%), suggesting a

Table 4.2 Comparative Performance of TIGs Across Datasets, with bolded values indicating the best performance per metric within each dataset, highlighting each tool's effectiveness across Detection Rate (DDR), Attack Success Rate (ASR), Perturbation Magnitude (PM), and Perceptual Similarity (LPIPS) in varying dataset complexities.

		RQ1: Robustness		**RQ2: Naturalness**	**RQ3: Diversity**		**RQ4: Efficiency**
Tool	**Dataset**	**DDR**	**ASR**	**LPIPS**	**PM (Avg)**	**PM (Std)**	**Time(s)**
Deephunter	MNIST	71.4%	38.24%	0.45	2.50	0.64	2.15
	CIFAR-10	72.04%	5.97%	**0.23**	**4.47**	1.81	46.34
	ImageNet-1k	34.0%	2.27%	**0.22**	12.26	**4.56**	1,080.00
Deepfault	MNIST	1.0%	1.00%	0.17	**0.13**	0.017	**0.178**
	CIFAR-10	**90.0%**	**90.0%**	0.61	7.85	0.51	**0.974**
	ImageNet-1k	42.6%	42.6%	0.44	5.79	0.3	743.35
AdvGAN	MNIST	**99.1%**	**99.1%**	**0.12**	3.83	0.52	15.60
	CIFAR-10	83.8%	83.8%	0.29	5.34	0.72	25.20
	ImageNet-1k	32.0%	32.0%	0.30	**1.28**	0.1	**60.00**
SinVAD	MNIST	89.0%	89.0%	0.29	9.41	**1.45**	57.60
	CIFAR-10	52.3%	80.2%	0.56	28.75	**4.93**	87.00
	ImageNet-1k	**100.0%**	**100.0%**	0.94	*inf*	0.0	540.00

All execution times are reported in seconds per image.

mismatch between its fault localization strategy and the feature-simple dataset.

For CIFAR-10, DeepFault excelled (DDR/ASR: 90.0%), effectively using targeted perturbations in a feature-rich dataset. AdvGAN performed well (83.8%), while SinVAD achieved a high ASR (80.2%) but a lower DDR (52.3%), indicating a focus on adversarial examples over defect detection. DeepHunter scored a low ASR (DDR: 72.04%, ASR: 5.97%), suggesting that its perturbation techniques may struggle with feature-rich datasets.

For ImageNet-1k, performance declined across all TIGs. While SinVAD achieved a DDR/ASR of 100%, it generated invalid test inputs (e.g., black images), highlighting an overfitting issue to simplistic patterns, which will be further discussed in RQ2. DeepFault showed consistent performance (DDR/ASR: 42.6%), while AdvGAN and DeepHunter achieved similar DDRs (32–34%), with DeepHunter's ASR being particularly low (2.27%), highlighting its ineffectiveness with high-resolution data.

Statistical analysis confirmed significant differences among TIGs (p-value < 0.05), with a large effect size ($\hat{A}_{12} > 0.5$) in most cases.

Findings: GMA TIGs (AdvGAN and SinVAD) excel on simpler datasets but face challenges with complex datasets. DeepFault performs consistently, especially on complex datasets. DeepHunter, constrained by its small perturbations, is less effective as dataset complexity increases. **Challenges:** SinVAD generated Out-of-Distribution inputs for ImageNet-1k, struggling to produce valid examples.

4.3.2 RQ2: Which TIG generates more natural test cases?

Motivation. The purpose of test cases is to identify DL erroneous behavior that could potentially occur in practice. Hence, assessing the naturalness of synthetically generated test cases is fundamental to infer their usefulness in real-world scenarios.

Method. We evaluate the naturalness of test cases using the Learned Perceptual Image Patch Similarity (LPIPS). LPIPS provides a perceptual similarity score between original and generated images, where lower values indicate more natural-looking images.

Results: The fifth column of Table 4.2 reports LPIPS findings, revealing that TIG naturalness is highly dataset-dependent and influenced by model architecture.

For MNIST, AdvGAN achieved the lowest LPIPS score (0.12), indicating that it generated the most natural-looking test cases. This aligns with GAN's ability to learn the underlying data distribution and produce realistic perturbations that are perceptually similar to the original inputs. DeepFault and SinVAD followed with LPIPS scores of 0.17 and 0.29, respectively, showing relatively natural perturbations. In contrast, DeepHunter had the highest LPIPS score (0.45), suggesting less natural test cases, likely due to more visible, pixel-level changes.

For CIFAR-10, DeepHunter produced the most natural test cases (LPIPS: 0.23), indicating that its perturbation strategy aligned well with the dataset's features. AdvGAN performed comparably (0.29), suggesting only minor perceptual differences. However, SinVAD and DeepFault had higher LPIPS scores (0.56 and 0.61), indicating a drop in naturalness. This decline can be attributed to the internal functioning of these TIGs: SinVAD relies on a VAE, which struggled with the rich features of CIFAR-10, leading to less realistic outputs. DeepFault, with its targeted neuron activation strategy, applied more aggressive modifications that disrupted the natural appearance of the test cases.

For ImageNet-1k, DeepHunter again led in naturalness (LPIPS: 0.22), followed by AdvGAN (0.30). DeepFault showed a degradation in naturalness (0.44), while SinVAD had the highest LPIPS score (0.94), indicating the least natural results. The poor performance of SinVAD on this dataset can be attributed to the VAE's difficulty in generating meaningful inputs,

as the generative model struggled to learn the complex features of ImageNet-1k, resulting in unrealistic outputs.

Overall, an inverse correlation between DDR/ASR and naturalness scores was observed. High DDR and ASR often correspond with lower naturalness, as more aggressive modifications are typically required to induce misclassifications, particularly in complex datasets.

Findings: GMAs (AdvGAN and SinVAD) tend to produce more natural modifications for simpler datasets, while PBAs like DeepHunter excel in generating natural-looking test cases for more complex datasets. An inverse relationship between robustness-revealing metrics (DDR/ASR) and naturalness was observed.

4.3.3 RQ3: Which TIG generates more diversified test cases?

Motivation. Diversified test cases are crucial to evaluate DL models in broader input space and under different situations. This RQ aims to assess each TIG's ability to generate varied test cases that explore different parts of the input space.

Method: We evaluate the diversity of the generated test cases using the Perturbation Magnitude (PM) metric, which quantifies the extent of modifications made to induce model misbehavior. Two key measures are considered: PM Avg, representing the average magnitude of perturbations applied to the original images, where higher values indicate stronger perturbations, and PM Standard deviation (Std), reflecting the variability in perturbation magnitude across the generated inputs, with higher values indicating greater diversity in the test cases.

Results: The sixth and seventh columns of Table 4.2 report Perturbation magnitude average (PM Avg) and Perturbation magnitude standard deviation (PM Std) findings, highlighting differences in perturbation strength and diversity across TIGs.

For MNIST, SinVAD had the highest PM Avg (9.41) and PM Std (1.45), reflecting strong and varied perturbations, making it the most diverse TIG. DeepHunter followed with a moderate PM Avg (2.50) and a high PM Std (0.64), indicating diverse but controlled perturbations. AdvGAN showed less diversity (PM Std: 0.52) but higher perturbation strength (PM Avg: 3.83). DeepFault exhibited minimal and uniform changes, with the lowest PM Avg (0.13) and PM Std (0.17).

For CIFAR-10, SinVAD again led in perturbation strength (PM Avg: 28.75) and diversity (PM Std: 4.93). DeepHunter achieved a balance with a moderate PM Avg (4.47) and higher PM Std (1.81). In contrast, AdvGAN (PM Avg: 5.34, PM Std: 0.72), and DeepFault (PM

Avg: 7.85, PM Std: 0.51) applied stronger but less varied perturbations.

For ImageNet-1k, SinVAD failed, producing extreme and uniform perturbations (PM Avg: ∞, PM Std: 0.0), generating out-of-distribution data. DeepHunter achieved a balanced performance (PM Avg: 12.26, PM Std: 4.56), with moderate perturbation strength and diversity. AdvGAN (PM Avg: 1.28, PM Std: 0.1) and DeepFault (PM Avg: 5.79, PM Std: 0.3) exhibited limited diversity and weaker perturbations, struggling with the complexity of ImageNet-1k.

Findings: SinVAD shows the highest diversity across simpler datasets but fails on ImageNet-1k. DeepHunter consistently balances perturbation strength and diversity across all datasets, making it the most reliable TIG. AdvGAN and DeepFault apply stronger but less diverse perturbations, indicating a focus on generating aggressive perturbations rather than exploring varied test cases.

4.3.4 RQ4: Which TIG is more efficient in test case generation?

Motivation. A TIG's efficiency is essential for promoting its adoption in real-world software testing scenarios. Efficient TIGs enable faster defect detection and optimize resource usage. This RQ evaluates the performance of each TIG while generating test inputs.

Method. We measured the execution time of each TIG across three datasets (MNIST, CIFAR-10, and ImageNet-1k). Each tool was run on 10 seed inputs, and the average execution time was calculated. To ensure comparability, execution times were normalized to a per-image basis, and all times were standardized to seconds.

Results: The eighth column of Table 4.2 reports the normalized execution times (in seconds per image) for each TIG across the three datasets.

DeepFault achieved the fastest execution on MNIST (0.178 s/image) and CIFAR-10 (0.974 s/image). However, its performance degraded significantly on ImageNet-1k (743.35 s/image), due to (1) the increased image resolution in ImageNet-1k (300×300) and (2) the computational demands when analyzing the more complex EfficientNet-B3 model. AdvGAN showed the highest scalability across all datasets, with relatively low execution times: 15.60 s/image for MNIST, 25.20 s/image for CIFAR-10, and only 60 s/image for ImageNet-1k. This consistent performance can be attributed to its generative model, which produces adversarial examples in a single forward pass without iterative optimization, making it highly efficient even on large datasets.

In contrast, SinVAD exhibited the highest execution times on smaller datasets (57.60 s/image

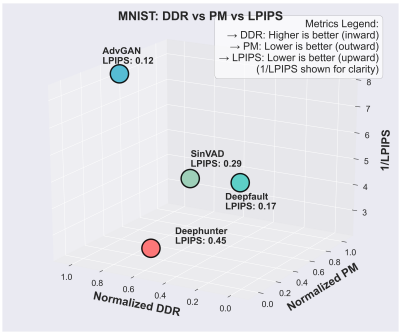
for MNIST and 87.00 s/image for CIFAR-10), primarily due to the overhead of training its VAE model. Despite this, SinVAD's execution time became more competitive for ImageNet-1k (540 s/image), reducing the relative impact of its initial overhead. DeepHunter showed competitive performance on smaller datasets, with execution times of 2.15 s/image for MNIST and 46.34 s/image for CIFAR-10. However, it exhibited a drastic increase to 1080 s/image on ImageNet-1k. This substantial increase can be explained by its reliance on coverage-guided fuzzing, which involves extensive iterative modifications and evaluation steps.

Findings: AdvGAN was the most efficient overall, with consistent performance across all datasets. DeepFault excelled on smaller datasets but faced scalability issues on ImageNet-1k. SinVAD's overhead affected its performance on small datasets but its overall performance improved on larger data. DeepHunter was efficient on simpler datasets but had significant delays on ImageNet-1k due to its iterative fuzzing approach.

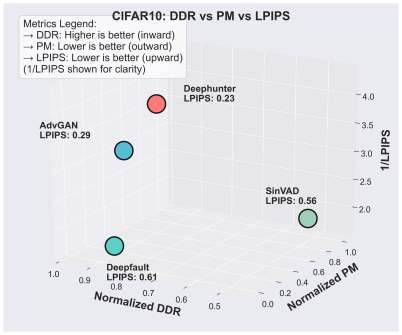
4.4 Discussion

Assessing TIGs: Strengths, Weaknesses, and Paths Forward

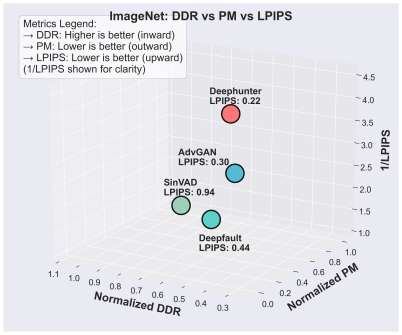
The evaluation of TIGs in our study highlights a range of distinct strengths, limitations, and inherent trade-offs among the different approaches 4.2. AdvGAN demonstrates a strong balance between revealing robustness issues and maintaining natural perturbations on simpler datasets. However, it struggles to scale effectively with high-resolution and complex data. Its strength lies in generating minimal perturbations, though this comes at the cost of reduced effectiveness on more complex datasets. Enhancing the generator and discriminator architectures could be a viable path to improving its efficacy on complex data. SinVAD, known for its ability to generate highly diverse test cases, also faces limitations when applied to high-resolution images due to the inherent limitations of its VAE architecture. The standard VAE employed in SinVAD lacks the expressive capacity to effectively model the complex data distributions found in high-resolution datasets. As a result, the generated samples often fall outside the valid data distribution, leading to unrealistic or out-of-distribution (OOD) test cases. To mitigate this issue, incorporating advanced generative models like Vector Quantized Variational Autoencoder (VQ-VAE) [104] could enhance the latent space representation. VQ-VAE, with its discrete latent codebook and improved expressiveness, may enable SinVAD to better capture the intricate data patterns of high-resolution images, thereby producing more realistic and in-distribution test cases.



(a) 3D Performance on MNIST



(b) 3D Performance on CIFAR-10



(c) 3D Performance on ImageNet

Figure 4.2 Comparison of 3D performance across different datasets. The points are small, so using high-resolution images and scaling them to the subfigure width helps keep them clear.

DeepFault and DeepHunter offer complementary strengths: DeepHunter excels in generating natural test cases, whereas DeepFault is more effective in exposing robustness issues. These tools show variable effectiveness across dataset complexities, underscoring a trade-off between robustness-revealing capability and maintaining natural perturbations. To address this trade-off, future work could focus on enhancing adaptive perturbation strategies, as well as leveraging recent advances in generative modeling.

Adaptation Challenges and Practical Considerations for TIGs

The process of adapting TIGs to different models and datasets revealed several practical challenges, particularly with newer, more complex architectures like EfficientNetB3. Many TIGs, including those evaluated in this study, were initially designed for simpler, fully connected models and older datasets. This legacy focus requires significant adaptation to accommodate modern architectures. Specifically, TIGs like DeepHunter rely on profiling methods tailored to earlier, shallower networks, which failed to work effectively with more recent models. Adjustments were required to enable these profiling techniques to function properly with current architectures.

Furthermore, the lack of modular design in many TIG codebases complicated the process of adapting them to new datasets. When switching from the default dataset, numerous changes were needed across various code files, making the adaptation process laborious. This experience highlights the importance of a modular architecture in TIGs, allowing users to easily reconfigure the system for different datasets and models without extensive code alterations. Finally, as illustrated in Figure 4.3, the performance of TIGs tends to decline

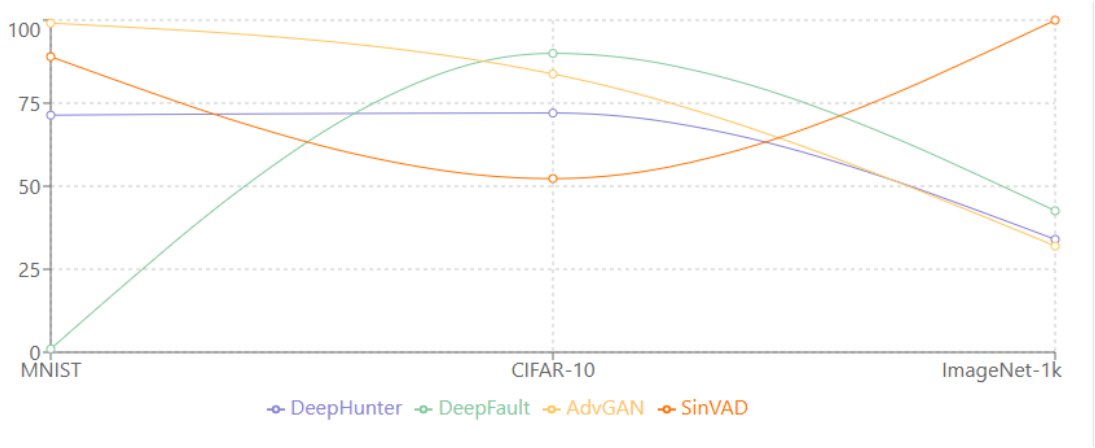


Figure 4.3 DDR Performance Trends of TIGs with Increasing Dataset Complexity, illustrating the scalability challenges encountered on more complex datasets.

as the complexity of the dataset increases. This observation suggests an urgent need for the development of more robust and adaptable TIGs that can effectively handle the demands of modern, complex model-dataset pairs.

More naturalness metrics are required

The relevance of a test case is directly proportional to the likelihood that the scenario it captures will occur in a real-world setting. Since DL systems are designed to handle real-world events, a test that induces erroneous behavior is only useful if it remains within realistic bounds; otherwise, its practical value is limited. Therefore, metrics and techniques for assessing and preserving the naturalness of generated test cases are crucial parts of every test generation approach's workflow. In this work, we used LPIPS as a proxy for naturalness. However, other naturalness measurements exist in the literature such as Inception Score [105] and Image Quality Assessment metrics [106, 107] and might provide contradictory findings since they portray naturalness from distinct perspectives. Measuring the visual quality of an image is very subjective, and there is no precise solution yet for formalizing its assessment, making naturalness a hot topic that requires further exploration. To address the limitations of automated metrics, researchers have used [108] human evaluations rather than these metrics to quantify naturalness. However, because using humans is not always feasible (e.g., in large-scale tasks), an automated approach must be developed.

Leveraging Large Language Models (LLMs) for Enhanced TIGs

LLMs offer promising opportunities for automating the generation of complex, diverse test inputs. Their generative power can efficiently produce edge cases and domain-specific inputs, reducing manual efforts and enhancing the realism of test cases, which may help identify vulnerabilities more effectively. Recent work, such as LANCE [109], has demonstrated the feasibility of using LLMs for generating test cases in image-based tasks. By leveraging large language modeling and text-based image editing, LANCE expands the range of test scenarios and highlights the potential of LLMs to enhance TIGs.

Looking ahead, LLMs could be utilized to create initial batches of diverse test cases, which traditional TIGs can then refine. This complementary strategy would leverage the strengths of both methods, increasing the diversity and complexity of test cases while minimizing computational overhead. Future work should explore deeper integration of LLMs to make TIGs more scalable and versatile, addressing a wider scope of robustness testing challenges.

4.5 Threats to Validity

In the following, we discuss the threats to the validity of our study.

Threats to internal validity. They may result from how the empirical study was conducted. To mitigate these issues, we used the default configurations of evaluated TIGs described in the original paper as the performance of TIGs can vary depending on the parameters selected. Additionally, to mitigate the risk of random variation affecting results, we repeated experiments 10 times. To confirm the statistical significance of our findings, we performed statistical hypothesis testing and effect size assessments using the non-parametric Wilcoxon test [101] and the Vargha–Delaney effect size [102].

Threats to external validity. They concern the applicability of our findings across different models, datasets, and TIGs. To mitigate these issues, we selected four state-of-the-art TIGs, three DL models of varying complexities, and three datasets of different sizes. This variety allowed us to evaluate the performance of TIGs in a broad spectrum of scenarios, from simple image datasets like MNIST to complex high-resolution datasets such as ImageNet-1K. However, our findings may not fully extend beyond image classification. Future work should assess TIGs in other domains, such as NLP or speech recognition, to enhance external validity.

Threats to conclusion validity. They involve the risk of bias arising from dataset and model selection and analysis procedures, potentially influencing the validity of the conclusions. To reduce selection bias, we chose well-known datasets (MNIST, CIFAR-10, ImageNet-

1K) and popular DL architectures (LeNet-5, VGG16, EfficientNetB3) commonly used in the DL and software testing communities.

4.6 Chapter summary

Our empirical study of four leading TIGs across datasets of varying complexity highlights that TIG performance varies considerably across diverse metrics, including robustness detection (DDR, ASR), naturalness (LPIPS), diversity (PM), and computational efficiency. For instance, while AdvGAN achieved exceptional performance on MNIST (DDR/ASR: 99.1%, LPIPS: 0.12), its effectiveness declined significantly on ImageNet-1k (DDR/ASR: 32.0%, LPIPS: 0.30). Similarly, DeepFault showed contrasting results between MNIST (DDR/ASR: 1.0%) and CIFAR-10 (DDR/ASR: 90.0%). Dataset complexity can also impact TIG performance. As dataset complexity increases, TIGs face growing challenges, including prolonged execution times and fluctuating effectiveness across the evaluation metrics. For instance, DeepHunter's execution time increased dramatically from 2.15s on MNIST to 1,080s per image on ImageNet-1k, while SinVAD's naturalness degraded significantly (LPIPS from 0.29 to 0.94) and ultimately failed to generate valid inputs for ImageNet-1k.

Our findings underscore the importance of selecting TIGs based on dataset characteristics, testing objectives, and resource constraints, highlighting the need for more adaptable, modular testing frameworks. Future work should focus on developing efficient, scalable methodologies that account for the growing complexity of DL systems and refining naturalness metrics to capture perceptually meaningful variations.

CHAPTER 5 EVALUATING AND ENHANCING SEGMENTATION MODEL ROBUSTNESS WITH METAMORPHIC TESTING

5.1 Chapter Overview

This chapter presents SegRMT, a novel approach that leverages metamorphic testing and genetic algorithms to evaluate and enhance segmentation model robustness. We implement SegRMT to generate adversarial examples that maintain visual coherence while effectively challenging model performance. Through experiments on the Cityscapes dataset with DeepLabV3, we demonstrate SegRMT's capabilities against traditional gradient-based methods. We explore adversarial training strategies and conduct cross-adversarial testing to assess generalization capabilities. The findings highlight SegRMT's potential for improving segmentation model reliability in unpredictable real-world environments where maintaining performance despite varying conditions is crucial.

5.2 Context of the Study

Despite the extensive literature about deep learning segmentation models—which aim to automatically partition images into distinct meaningful regions— and their applications, little attention has been given to their robustness against adversarial attacks or image distortions so far. Nowadays, with the rise of deep learning segmentation models, there is a growing demand for evaluating their robustness, particularly in critical fields like autonomous driving and medical imaging [110, 111]. These critical applications require highly reliable models to prevent potentially devastating consequences. For example, accurately segmenting road scenes under diverse conditions in autonomous driving is crucial for ensuring safety. In medical imaging, for example, it is equally important that the identification and segmentation of the anatomical structures are as accurate as necessary to ensure correct diagnosis and effective treatment planning. Recent studies suggest that adversarial data, whether controlled (e.g., human-induced attacks) or uncontrolled (e.g., distortions), are on the rise and are hindering the robustness of deep learning segmentation models [112, 113]. Adversarial human-induced attacks involve introducing artificial alterations to input data in order to deceive the models, resulting in inaccurate predictions and compromising their reliability. Conversely, distortions induced by adversarial environments involve uncontrollable factors, such as variations in lighting, occlusions, and sensor noise, which can also impact the model's performance and compromise their reliability. When testing the robustness of models in real-world applica-

tions, traditional gradient methods like the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) are commonly used. However, these methods may not fully account for other adversaries, such as distortions caused by adversarial environments [114–116].

Recently, Metamorphic testing (MT) has emerged as a promising approach to combine gradient-based adversarial attacks and real-world distortions to evaluate and enhance models robustness [113]. MT simulates real-world disturbances and generates new test cases by systematically modifying existing ones to evaluate model robustness under diverse input conditions, incorporating adversarial attacks and real-world disturbances [117–119]. When incorporated with an optimization algorithm, MT provides scalability and optimization advantages to tune the distortions while avoiding corrupting the input data [113, 120, 121]. Therefore, in this work, we propose a novel optimization-driven approach to segmentation robustness testing. Our key innovation lies in combining metamorphic testing (MT) with genetic algorithms (GA) to intelligently discover and optimize adversarial distortions. Unlike existing approaches that use fixed transformation patterns or random perturbations, our segmentation robustness metamorphic testing approach (SegRMT) leverages GA's evolutionary optimization to systematically explore the space of possible transformations, identifying the most effective combinations of distortions while maintaining image fidelity. Furthermore, we introduce a threshold for Peak Signal-to-Noise Ratio (PSNR), a quality metric that measures the ratio between maximum possible pixel value and distortion noise, to ensure that SegRMT generates adversarial input data within safe limits for distortions and data integrity [120,122]. Otherwise, over-distortion may result in high corruption and diverge the input data away from the norms. We use the cityscapes dataset to evaluate and enhance the robustness of segmentation models across various experiments [123]. In the first experiment, we demonstrate the effectiveness of SegRMT compared to the traditional gradient-based adversarial attacks. In the second experiment, we fine-tune the segmentation model through adversarial training using training data and adversarial examples generated by SegRMT and traditional gradient-based adversarial attacks. This exposes the model to a wide range of perturbations during training, thus potentially improving its robustness under adversarial conditions and increasing its performance. In the third experiment, we cross-test the robustness of segmentation models between SegRMT and traditional gradient-based adversarial attacks.

The contribution of this work is twofold and includes: 1) proposing a novel GA-optimized framework for evaluating segmentation model robustness, which systematically discovers effective adversarial transformations through evolutionary optimization, and 2) enhancing the robustness of the model using adversarial training data. We express and validate these contributions by answering the following research questions.:

- **RQ1** How effective is SegRMT compared to traditional gradient-based adversarial attacks in deceiving a segmentation model?
- **RQ2** How effectively does SegRMT enhance segmentation model robustness in self-adversarial testing compared to traditional gradient-based attacks?
- **RQ3** How effectively does SegRMT improve segmentation model robustness in cross-adversarial testing versus traditional gradient-based attacks?

Importantly, while previous work has combined metamorphic testing with genetic algorithms for applications in software testing [124,125] and classification models robustness testing [113], these approaches have not been extended to image segmentation. Segmentation models operate on high-dimensional, pixel-level outputs and require strict preservation of both visual and semantic integrity—challenges that are unique to this domain. Our approach, therefore, represents the first application of metamorphic testing with GA to assess and enhance segmentation robustness. This novel application not only tailors the optimization process to address the specific challenges of segmentation but also demonstrates superior performance against traditional gradient-based adversarial attacks. Our first finding shows that traditional gradient-based adversarial attacks decrease the mean Intersection over Union (mIoU), a metric that quantifies segmentation accuracy by measuring overlap between predicted and ground truth regions, to a minimum of 8.5% at a PSNR of 21.8 dB. In comparison, SegRMT can decrease the mIoU even further to a minimum of 6.4% at a higher PSNR of 24 dB. This indicates that SegRMT often generates more challenging adversarial examples than traditional gradient-based attacks. The second finding demonstrates that SegRMT enhances the robustness of models fine-tuned by self-adversarial testing, but it may not surpass traditional gradient-based adversarial attacks. The third finding shows that the model fine-tuned on SegRMT adversarial achieves a maximum of 68.0% as mIoU against the traditional gradient-based adversarial examples, while other models fine-tuned on traditional gradient-based adversarial achieve a maximum of 10.0% as mIoU against SegRMT adversarial examples.

The remainder of the chapter is organized as follows. Section 5.3 formulates the problem and identifies our research's objectives, constraints, and variables. Section 5.4 describes the proposed approach, including the design and implementation of the metamorphic testing and adversarial training methods. Section 5.5 illustrates our experimental setup and reports our findings on evaluating and enhancing the robustness of segmentation models. Finally, Section 5.8 concludes the chapter.

5.3 Problem formulation

In this section, we define and structure the research problem. Also, we identify the objectives, constraints, problem specifications, and variables involved in our research.

In this study, we rigorously investigate the robustness of segmentation models tailored to diverse imaging modalities, encompassing hyperspectral, multispectral, and standard RGB datasets. Our dataset, D, is formally defined as: $D = \{(X_i, Y_i)\}_{i=1}^{N}$, where each X_i represents an image instance from the aforementioned modalities and Y_i denotes its corresponding accurate ground truth segmentation map. The primary objective for our segmentation model F is the precise mapping of each image X_i to its expected segmentation output, aiming for a high fidelity approximation $F(X_i) \approx Y_i$.

To methodically assess and enhance the model's resilience against various input perturbations, we employ an array of metamorphic relations R. These relations represent systematic ways to transform images while preserving their essential characteristics. For instance, in an autonomous driving context, if we slightly adjust an image's brightness (simulating different times of day) or add minor noise (simulating sensor interference), a car should still be recognized as a car in the segmentation output. Each metamorphic relation $r \in R$ defines a deliberate, parameterized transformation T_r, conceived to simulate potential real-world alterations affecting the images: $T_r(X_i, \theta_r) = \hat{X}_i$, where θ_r encapsulates the parameters of the transformation, constrained within a defined permissible range Θ.

The core premise of our metamorphic testing protocol insists that, despite these transformations, the essential semantic integrity of the image segments must be preserved, i.e., $F(T_r(X_i, \theta_r)) \approx Y_i$. This condition forms the basis for asserting the robustness of our model, ensuring that the semantic content of the segments remains intact despite the application of T_r.

We approach this challenge by formulating a constrained optimization problem designed to maximize a robustness criterion $C(\hat{X})$, which quantitatively evaluates the model's ability to uphold segmentation accuracy in the face of these synthetic perturbations. This is mathematically expressed as:

$$\text{Maximize } C(\hat{X}) \text{ subject to } \Phi_i(\hat{X}) = 0, \text{ for } i = 1, \ldots, u$$

Here, Φ represents a set of validity constraints ensuring that the transformations T_r respect the semantic boundaries as defined by the original ground truths. Specifically, our validity constraints consist of a Peak Signal-to-Noise Ratio (PSNR) threshold of 20dB and realistic transformation requirements. The PSNR threshold ensures that our transformations

maintain sufficient image quality and visual coherence while still allowing for meaningful perturbations. The realism constraints guarantee that our transformations simulate real-world scenarios that could naturally occur during image acquisition and processing, such as lighting variations, sensor noise, or perspective changes. Together, these constraints ensure that our adversarial examples remain both challenging and representative of real-world conditions that a segmentation model might encounter during deployment.

To augment the model's resilience further, adversarially generated examples D_{adv} — synthetic inputs that significantly deviate from expected outcomes under nominal conditions — are integrated into the training dataset D_{train}:

$$D_{\text{augm}} = D_{\text{train}} \cup D_{\text{adv}}$$

Subsequently, the model F undergoes fine-tuning on D_{augm} aimed at minimizing the empirical error, measured via a loss function L, across both original and adversarial examples:

$$\text{Minimize } \mathbb{E}[(X, Y) \sim D_{\text{augm}}][L(F(X), Y)]$$

This meticulous approach, employing metamorphic testing and adversarial training, not only fortifies the segmentation models against a spectrum of challenging conditions but also systematically refines their accuracy and reliability across varied imaging contexts. This methodology ensures a comprehensive evaluation and continuous enhancement of the models' segmentation capabilities, embodying a robust defense against real-world perturbations and synthetic adversarial tactics.

5.4 Methodology

In this section, we describe the use of metamorphic testing combined with the PSNR constraints. Also, we explain how we optimize image perturbations using the genetic algorithm.

We employ a metamorphic testing framework to evaluate our segmentation model's robustness thoroughly and rigorously. This framework systematically applies a series of controlled transformations to the input images, known as metamorphic relations. These transformations are designed to simulate a variety of real-world perturbations and variations, thereby providing a comprehensive assessment of the model's performance under challenging conditions.

Figure 5.1 shows the overall process of generating and evaluating transformation vectors, which are crucial in SegRMT. The process consists of two main components highlighted in

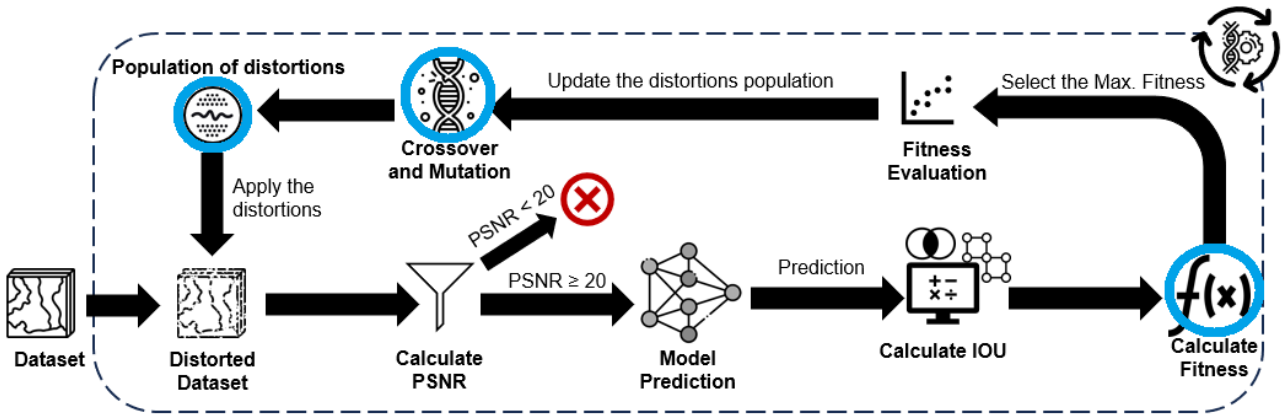


Figure 5.1 Pipeline of the proposed SegRMT for robustness assessment. The pipeline illustrates the process from initial image perturbation using various transformations and optimization using the genetic algorithm to the evaluation of segmentation model performance.

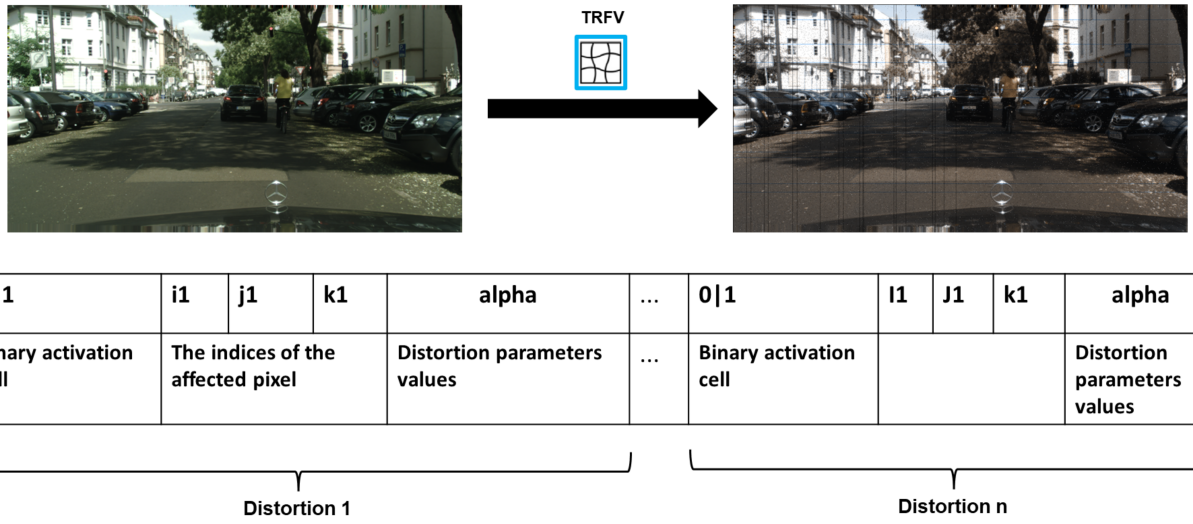


Figure 5.2 Transformation vector structure.

Figure 1. The first component is the transformation vector generation, where vectors are randomly initialized. Each vector consists of sub-transformation vectors representing specific types of noise or perturbation. As shown in Figure 2, each sub-transformation vector includes 1) a binary activation cell indicating whether the transformation is active, 2) indices of the affected pixels, and 3) distortion parameter values. The second component is the genetic algorithm optimization, which evaluates and evolves these transformation vectors to find the most effective perturbations. The GA employs a carefully designed fitness function that balances two objectives: maximizing the PSNR to ensure image quality while minimizing the Intersection over Union (IoU) to identify transformations that significantly impact

model performance. Through this optimization process, the GA aims to discover the least perceptible transformations that can effectively challenge the segmentation model.

5.4.1 Image Transformation

In real-world deployments, image segmentation models must handle various forms of image degradation. Our transformation selection methodology addresses these challenges through two fundamental categories: spatial and spectral distortions, reflecting the primary ways images can be compromised in practical applications. This categorization is motivated by both empirical studies and practical considerations. For example, Hendrycks et al. [126]demonstrated that a wide range of noise, blur, and digital corruptions—similar to those we consider—can significantly affect model performance, while Taori et al. [127] highlighted the importance of testing models against natural distribution shifts using diverse perturbations. Building on these insights, we group the most critical real-world corruptions into spatial distortions, which simulate physical and geometric alterations to the image structure, and spectral distortions, which target color information processing. Specifically, we focus on four key spatial distortions—region dropout (occlusions/missing data), line/column transformations (systematic sensor errors), salt and pepper noise (impulse distortions), and Gaussian noise (thermal variations)—and two spectral distortions—channel dropout and channel-specific noise—representing color degradation scenarios common in real imaging systems.

Spatial Distortions

Let $I : \Omega \to \mathbb{R}^C$ be an image defined on the pixel grid $\Omega \subset \mathbb{Z}^2$, where C is the number of channels (e.g., $C = 3$ for RGB images). For clarity, we define:

- $\text{MIN}_I = \min_{(x,y)\in\Omega} I(x, y)$,
- $\text{MAX}_I = \max_{(x,y)\in\Omega} I(x, y)$,
- CONST_I: a predetermined constant value (typically chosen as either MIN_I or MAX_I).

Spatial distortions are applied to test the model's ability to handle changes in the spatial domain. These distortions include:

- **Region Dropout:**
  - **Description:** Simulates occlusions or missing data by randomly altering regions in the image.

- **Purpose:** This transformation mimics real-world scenarios where parts of an image might be blocked by objects, sensor malfunctions, or environmental obstructions. In segmentation tasks, occlusions are common—such as pedestrians partially hidden behind vehicles or objects obscured by shadows. By applying region dropout, the model is forced to rely on contextual cues from the remaining visible parts, thereby testing its ability to infer and preserve semantic information even when significant regions are missing.
- **Mathematical Formulation:** For each pixel $(x, y) \in \Omega$, define:

$$I'(x,y) = \begin{cases} \text{MIN}_I, & \text{with probability } p_{\text{min}}, \\ I(x,y), & \text{with probability } p_{\text{unchanged}}, \\ \text{MAX}_I, & \text{with probability } p_{\text{max}}, \end{cases}$$

where $p_{\text{min}} + p_{\text{unchanged}} + p_{\text{max}} = 1$.

- **Line and Column Transformations:**
  - **Description:** Modifies entire rows or columns of pixels to simulate sensor errors or calibration issues. Such transformations have been shown to effectively model structured perturbations.
  - **Purpose:** These transformations represent systematic, structured distortions that can occur due to hardware issues (e.g., faulty sensor lines) or calibration errors in imaging devices. Such errors can lead to consistent distortions across an image. For segmentation, where continuity and precise boundaries are critical, these structured perturbations test the model's resilience against uniform or patterned noise, ensuring that it can still accurately delineate object boundaries despite consistent, directionally biased distortions.
  - **Continuous Line/Column Dropout:** Let ℓ denote a specific row or column index. Then, for all $(x, y) \in \Omega$, one can set:

$$I'(\ell, y) = \text{CONST}_I \quad \text{or} \quad I'(x, \ell) = \text{CONST}_I.$$

  - **Line Stripping:** For a given stride $s \in \mathbb{N}$, define:

$$I'(x,y) = \begin{cases} I(x,y), & \text{if } x \bmod s \neq 0, \\ \text{CONST}_I, & \text{if } x \bmod s = 0, \end{cases}$$

where x mod s denotes the remainder when x is divided by s. A similar formulation applies for column stripping.

- **Salt and Pepper Noise:**
  - **Description:** Introduces random occurrences of black and white pixels.
  - **Purpose:** Salt and pepper noise is a classic model for impulse noise, often arising from errors in data transmission or sensor defects. In practical imaging scenarios, sudden and isolated pixel-level disturbances may occur due to environmental interference or hardware glitches. For segmentation models, handling such abrupt changes without losing overall structural information is crucial. This transformation challenges the model to remain robust in the presence of isolated, high-contrast pixel anomalies.
  - **Mathematical Formulation:** For each pixel $(x, y) \in \Omega$,

$$I'(x,y) = \begin{cases} \text{MIN}_I, & \text{with probability } p_{\text{salt}}, \\ I(x,y), & \text{with probability } 1 - \Big(p_{\text{salt}} + p_{\text{pepper}}\Big), \\ \text{MAX}_I, & \text{with probability } p_{\text{pepper}}, \end{cases}$$

    with $p_{\text{salt}} + p_{\text{pepper}} \leq 1$.

- **Spatial Gaussian Noise:**
  - **Description:** Adds Gaussian-distributed noise to the pixel values.
  - **Purpose:** Gaussian noise represents natural fluctuations that occur in sensor readings (e.g., thermal noise). Unlike impulse noise, Gaussian noise is spread throughout the image and tends to be less abrupt, but it still affects the clarity of edges and textures. By introducing spatial Gaussian noise, the model is tested on its ability to distinguish important structural features from the inherent noise present in real-world imaging, ensuring that minor variations in pixel intensities do not lead to significant mis-segmentation.
  - **Mathematical Formulation:**

$$I'(x,y) = I(x,y) + \eta(x,y),$$

    where $\eta(x,y) \sim N(\mu, \sigma^2)$ are independent samples drawn from a Gaussian distribution with mean μ and variance σ^2, for each $(x,y) \in \Omega$.

Spectral Distortions

Spectral distortions test the model's robustness to variations in color channels. Although these are typically applied to multi-spectral images, they can be adapted to RGB images.

- **Channel Dropout:**
  - **Description:** Simulates the loss of specific color channels in RGB images.
  - **Purpose:** In real-world conditions, sensors may sometimes fail to capture complete color information due to hardware faults or adverse lighting conditions. Channel dropout forces the segmentation model to operate with incomplete color information, testing its robustness to missing data. This is especially critical for segmentation tasks where color cues often play a significant role in differentiating between objects with similar shapes but different colors.
  - **Mathematical Formulation:**

$$I'_c(x, y) = \text{CONST}_I,$$

    where $c \in \{R, G, B\}$ denotes a specific color channel.

- **Channel Gaussian Noise:**
  - **Description:** Introduces Gaussian noise to specific color channels in RGB images.
  - **Purpose:**This transformation addresses situations where one or more color channels might exhibit slight, channel-specific variations due to environmental changes or sensor inconsistencies. Since color fidelity is important for accurately distinguishing between objects—especially in scenarios where objects have similar textures but differing color profiles—this transformation ensures that the model can handle slight fluctuations in individual channels without degrading segmentation performance. It also reinforces the idea that the model should not be overly sensitive to minor color variations, which are common in real-world settings. [128].
  - **Mathematical Formulation:**

$$I'_c(x, y) = I_c(x, y) + \eta_c(x, y),$$

    where $\eta_c(x, y) \sim N(\mu, \sigma^2)$ represents the noise added to channel c.

Together, these transformations capture many common real-world degradations. We selected region dropout, line/column transformations, salt and pepper noise, and spatial Gaussian noise for the spatial domain because they effectively simulate issues such as occlusions, sensor malfunctions, and natural noise variations—conditions frequently encountered in applications like autonomous driving. Similarly, the spectral perturbations (channel dropout and channel-specific Gaussian noise) mimic failures in color capture due to hardware faults or adverse lighting conditions. These choices are grounded in established research [126–129], ensuring that our transformation set is both comprehensive and relevant to practical scenarios. While this set robustly challenges segmentation models, future work may explore additional transformations (e.g., geometric rotations, scaling, and weather-induced effects) to further expand the evaluation of model robustness.

5.4.2 Robustness Criterion

Let $F(\hat{X}_i)$ denote the segmentation output for the perturbed image $\hat{X}_i$, and let Y_i be the corresponding ground truth segmentation map. The robustness criterion, which quantifies the model's ability to maintain segmentation accuracy under perturbations, is defined as:

$$C(\hat{X}) = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}\Big(F(\hat{X}_i), Y_i\Big),$$

where IoU (Intersection over Union) measures the overlap between the predicted segmentation $F(\hat{X}_i)$ and the ground truth Y_i. A higher $C(\hat{X})$ value indicates better robustness.

5.4.3 Genetic Algorithm for Optimizing Transformations

The pursuit of reliable segmentation models led to the identification of a notable gap in current approaches related to the generation of adversarial examples. Although fixed transformations and random perturbations have their advantages, adopting a more advanced methodology has the potential to provide adversarial examples that are both more realistic and sophisticated. In light of this revelation, the investigation of Genetic Algorithms (GAs) in this particular field was initiated.

Genetic Algorithms (GAs), which are based on principles from evolutionary biology, provide a promising foundation for effectively navigating the intricate search space of image transformations. The underlying logic was that by emulating the process of natural selection, it could be feasible to develop progressively more efficient combinations of transformations, thus expanding the limits of what conventional approaches could achieve. Following a comprehen-

sive examination of relevant academic articles and careful consideration, the determination was reached to employ a customized Genetic Algorithm (GA) specifically designed to address the unique requirements of this study. The choice to enhance the complexity of the project was considered appropriate due to the potential benefits it could bring in terms of improving the quality and diversity of adversarial examples.

The design of the chromosome structure is a critical component of the genetic algorithm, directly influencing its ability to represent and evolve effective transformations. After careful consideration and multiple iterations, a sophisticated chromosome structure was developed to encode complex sequences of transformations.

Each chromosome consists of several sub-transformation vectors, each representing a specific type of distortion that can be applied to the input image. Figure 5.2 shows the structure of each sub-transformation vector is as follows:

- **Binary Activation Cell**: A single bit (0 or 1) indicating whether this particular distortion should be applied.
- **Distortion Parameters**: A set of values specific to the type of distortion (e.g., dropout rate, noise variance, color shift values).
- **Affected Indices**: Specifies which pixels or bands the distortion should be applied to.

This detailed encoding ensures a deterministic mapping between the chromosome and the resulting distorted input, given the original image. It allows for fine-grained control over the application of distortions while maintaining the flexibility to represent a wide variety of transformation combinations.

The population initialization process was carefully designed to generate a diverse set of valid chromosomes. This involved:

1. Randomly determining the number of sub-transformation vectors for each chromosome.
2. For each sub-transformation vector:
   (a) Randomly setting the activation bit.
   (b) Generating appropriate parameter values within predefined ranges specific to each distortion type.
   (c) Selecting affected indices based on the distortion type and image dimensions.

A validation step was implemented to ensure that all initial chromosomes represented feasible transformation sequences. This extra layer of validation significantly reduced errors in subsequent generations and ensured that the genetic algorithm started with a population of viable solutions.

This chromosome design, coupled with the carefully crafted initialization process, provided a solid foundation for the genetic algorithm to explore and evolve increasingly effective combinations of image transformations.

A key innovation in our GA implementation is the development of a sophisticated fitness function that balances two competing objectives: maximizing the disruption of segmentation results while preserving image fidelity. After extensive experimentation, we developed the following formulation:

$$F = \begin{cases} (1 - \text{IoU}) \times \left(\frac{\text{PSNR}}{20}\right) & \text{if PSNR} \geq 20 \text{ dB} \\ 0 & \text{if PSNR} < 20 \text{ dB} \end{cases}$$

This formulation incorporates several key insights gained through the research process:

- **PSNR Normalization**: By dividing PSNR by the threshold value of 20 dB, the function creates a balanced interplay between segmentation disruption (measured by IoU) and image fidelity (measured by PSNR). This normalization ensures that neither objective dominates the fitness calculation.
- **IoU Inversion**: The use of $(1 - \text{IoU})$ in the formula ensures that lower IoU values, which indicate greater segmentation disruption, result in higher fitness scores. This aligns the fitness function with the goal of finding transformations that significantly impact segmentation performance.
- **Quality Threshold**: The implementation of a hard cutoff at 20 dB PSNR serves to eliminate transformations that excessively degrade image quality. This threshold was determined through a combination of literature review [113] and empirical testing, representing a balance point between perceptible image degradation and effective adversarial perturbation.

Through comprehensive testing across diverse transformation scenarios, this fitness function has demonstrated consistent ability to guide the GA towards transformations that are both subtle in visual impact and effective in disrupting segmentation performance. The combination of our sophisticated chromosome structure and carefully crafted fitness function provides

a solid foundation for the GA to explore and evolve increasingly effective combinations of image transformations.

5.5 Experiments

In this section, we will detail the Experimental setup used to conduct this study we will also detail and discuss the different results obtained offering insights into their significance and implications

5.5.1 Experimental Setup

Datasets: We used the Cityscapes dataset for our experiments, which consists of 5,000 high-resolution urban street images with a resolution of 2048x1024 pixels. The dataset is divided into 2,975 training images, 500 validation images, and 1,525 testing images. This dataset was selected due to its complexity and relevance to autonomous driving applications, providing a challenging environment for testing model robustness.

Models: Our experiments utilized the DeepLabV3 model with a ResNet-50 backbone, a well-established architecture for segmentation tasks. The model was pre-trained on the Cityscapes dataset, serving as a strong baseline for evaluating the impact of adversarial perturbations.

Baseline Methods: To generate adversarial examples, we implemented three widely recognized methods:

- **Fast Gradient Sign Method (FGSM):** Implemented with an epsilon value of 0.09 to achieve a PSNR value of approximately 20 dB.
- **Projected Gradient Descent (PGD):** Conducted with 10, 40, and 100 iterations using alpha and epsilon values of 0.08 and 0.09, respectively.
- **Carlini & Wagner (C&W) Attack:** Executed as an optimization problem with an epsilon value of 0.15 and a learning rate of 1e-5 to minimize perturbation while maintaining adversarial efficacy and a PSNR abover 20 dB.

Evaluation Metrics:
The effectiveness of the adversarial attacks was measured using two key metrics:

- **Intersection over Union (IoU):** This metric assesses the overlap between the predicted segmentation and the ground truth, indicating the accuracy of the segmentation.

- **Peak Signal-to-Noise Ratio (PSNR):** This metric quantifies the perceptual similarity between the original and adversarial images, ensuring that the adversarial examples maintain the semantic integrity of the images.

Genetic Algorithm Configuration: Through extensive experimentation, we established the following parameters for our genetic algorithm: a population size of 50 individuals, evolution limit of 100 generations, crossover rate of 0.8, and mutation rate of 0.2 with variable sub-rates for different mutation types. We preserved the top 2 individuals through elitism and implemented early termination when fitness improvement remained below 0.1% for 15 consecutive generations. This configuration provided an effective balance between exploration and exploitation.

Statistical Considerations:
To ensure the robustness of the results and enable statistical analysis, each experiment was repeated 10 times using different random seeds. The results from these multiple runs were used to perform statistical significance tests, such as the Wilcoxon signed-rank test, and calculate effect sizes (e.g., Cohen's d). This allowed for a more reliable evaluation of the differences between the SegRMT method and traditional gradient-based adversarial attacks.

5.5.2 Evaluating Segmentation Robustness

To assess our segmentation model's robustness, we conducted comparative experiments between our SegRMT approach and traditional adversarial methods (FGSM, PGD, and C&W attacks). Each method was calibrated to maintain a PSNR above 20 dB, ensuring fair comparison while preserving essential image content and semantics. Our evaluation used the DeepLabV3 model trained on the Cityscapes dataset, focusing on both original and adversarially perturbed images. The results, presented in Table 1, demonstrate SegRMT's superior effectiveness in generating challenging adversarial examples. Our approach achieved a significant reduction in model performance, lowering the mIoU to 6.4% while maintaining a higher PSNR of 24.0 dB. In comparison, traditional methods showed less effectiveness: FGSM reduced mIoU to 11.3% (PSNR 20.6 dB), PGD variants achieved between 8.5% and 9.4% (PSNR 21.8 dB), and C&W reached 21.7% (PSNR 21.3 dB). These results indicate that SegRMT generates more potent adversarial examples while better preserving image quality. Statistical analysis of our results, based on 10 repeated experiments with different random seeds, confirmed the significance of these findings. The Wilcoxon signed-rank test and Cohen's d effect size calculations demonstrated that SegRMT's performance improvements over traditional methods were both statistically significant and practically meaningful. This comprehensive evaluation framework revealed that our approach provides a more prac-

tical assessment of model robustness, particularly in scenarios requiring realistic perturbation patterns.

RQ1: How effective is SegRMT compared to traditional gradient-based adversarial attacks in deceiving a segmentation model?

Motivation: This research question is motivated by the necessity to thoroughly assess and improve the resilience of segmentation models, especially in scenarios where adversarial attacks might significantly hinder model performance. Conventional gradient-based adversarial techniques, including FGSM, PGD, and C&W, have been extensively used to stress-test the model used. However, they may not comprehensively capture the wide range of possible adversarial scenarios. By incorporating Metamorphic Testing (MT) and Genetic Algorithms (GA), SegRMT aims to investigate a broader spectrum of adversarial examples, potentially leading to more effective results.

Approach: To investigate the effectiveness of SegRMT, we subjected the DeepLabV3 model, trained on the Cityscapes dataset, to a comprehensive evaluation against various adversarial attacks. The resilience of the model was measured using two main metrics: mean Intersection over Union (mIoU) for segmentation accuracy and Peak Signal-to-Noise Ratio (PSNR) for evaluating the subtlety of the applied disturbances.

SegRMT, which integrates a Genetic Algorithm within a Metamorphic Testing framework, was systematically compared to traditional gradient-based adversarial methods, including FGSM, PGD, and C&W. The Genetic Algorithm (GA) in SegRMT optimizes perturbations by balancing the trade-off between minimizing mIoU and maximizing PSNR. A threshold of 20 dB is established to ensure that the adversarial examples remain perceptually realistic. This method enabled a meticulous evaluation of the effectiveness of SegRMT in generating challenging yet visually subtle adversarial examples.

Results: The results of the robustness testing, as shown in Table 5.1, demonstrate the effectiveness of the SegRMT approach compared to traditional gradient-based adversarial attacks on the DeepLabV3 model trained with the Cityscapes dataset. The model's initial performance on the unaltered dataset revealed a high mean Intersection over Union (mIoU) of 79.4%, highlighting its precision in optimal circumstances.

However, the model's ability to withstand adversarial perturbations significantly decreased, particularly with the SegRMT approach. The FGSM attack decreased the mean Intersec-

Table 5.1 Robustness Testing Results on Cityscapes with DeepLabV3

Testing Method	**mIoU** (%)	**PSNR (dB)**
Original (No Perturbation)	79.4	Inf
FGSM	11.3	20.6
PGD10	9.4	21.8
PGD40	8.5	21.8
PGD100	9.1	21.8
C&W	21.7	21.3
SegRMT	**6.4**	**24.0**

tion over Union (mIoU) to 11.3%, with a Peak Signal-to-Noise Ratio (PSNR) of 20.6 dB, underscoring the model's vulnerability to even basic gradient-based attacks. The iterative PGD method showed varying degrees of impact, with mIoU values ranging from 9.4% to 8.5% as iterations increased, though effectiveness plateaued beyond 40 iterations, indicating a diminishing return on computational effort.

The C&W attack, known for its accuracy, yielded a mean Intersection over Union (mIoU) of 21.7% and a Peak Signal-to-Noise Ratio (PSNR) of 21.3 dB. Nevertheless, SegRMT showed the most notable results by achieving the lowest mIoU of 6.4% while preserving the highest PSNR of 24 dB. This suggests that the use of a Genetic Algorithm in SegRMT's Metamorphic Testing framework enables the generation of highly effective adversarial instances that are both subtle and significantly impactful. These findings emphasize SegRMT's exceptional capacity to deteriorate model performance, making it a robust tool for evaluating the resilience of segmentation models against a broader range of adversarial scenarios.

To assess the statistical significance of the differences in performance between the metamorphic testing tool and the gradient-based adversarial attack methods, we conducted the Wilcoxon signed-rank test and calculated Cohen's d effect size. The Wilcoxon signed-rank test is a non-parametric statistical test used to compare two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. We performed pairwise comparisons between the metamorphic testing tool and each of the gradient-based methods (PGD10, PGD40, PGD100, CW, and FGSM) using the image-wise IoU scores. The results are as follows:

- Wilcoxon Test for SegRMT vs. PGD10: Test Statistic: 48918.0, P-value: 0.0047
- Wilcoxon Test for SegRMT vs. PGD40: Test Statistic: 38221.0, P-value: $2.099e^{-10}$
- Wilcoxon Test for SegRMT vs. PGD100: Test Statistic: 34520.0, P-value: $3.593e^{-14}$

- Wilcoxon Test for SegRMT vs. CW: Test Statistic: 830.0, P-value: $5.978e^{-78}$
- Wilcoxon Test for SegRMT vs. FGSM: Test Statistic: 16726.0, P-value: $3.270e^{-41}$

The low p-values (< 0.05) for all comparisons indicate that the differences in performance between the metamorphic testing tool and each of the gradient-based methods are statistically significant. This suggests that the tool's effectiveness in reducing the model's IoU scores is not due to chance.
To further quantify the magnitude of the difference between the metamorphic testing tool and the gradient-based methods, we calculated Cohen's d effect size. We separated the methods into two groups: the metamorphic testing tool in one group and the gradient-based methods (PGD10, PGD40, PGD100, CW, and FGSM) in another group. The result is as follows:

- Cohen's d for Tool vs. Gradient-Based Attacks: -0.641

The negative value of Cohen's d indicates that the metamorphic testing tool group has lower IoU scores than the gradient-based methods group. The absolute value of 0.641 suggests a medium to large effect size, indicating that the difference between the two groups is substantial and practically significant.
In summary, the statistical analysis using the Wilcoxon signed-rank test and Cohen's d effect size provides strong evidence that the metamorphic testing tool is more effective than the gradient-based adversarial attack methods in reducing the model's IoU scores. The differences are both statistically significant and practically meaningful, highlighting the tool's potential for robustness testing of deep learning models in semantic segmentation tasks.

5.5.3 Enhancing Robustness through Adversarial Training

This section details the specific experimental configurations and implementation details used in our adversarial training evaluations.

Implementation Configuration We conducted our experiments using the DeepLabV3 model with a ResNet-50 backbone on the Cityscapes dataset. Each experiment was performed using one of five distinct configurations:

1. Base model trained exclusively on clean data (baseline)
2. Model trained with FGSM adversarial augmentation
3. Model trained with PGD adversarial augmentation

4. Model trained with C&W adversarial augmentation
5. Model trained with SegRMT adversarial augmentation

Training Parameters For all configurations, we employed a Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. Given the high-resolution nature of Cityscapes images (2048x1024 pixels) and GPU memory constraints, we implemented a batch size of 2. All adversarial examples were generated while maintaining a minimum PSNR threshold of 20 dB to ensure data integrity.

Data Processing Our implementation included standard data augmentation techniques: random resizing, cropping to 512x1024 pixels, horizontal flipping, and photometric distortion. For adversarial configurations, we maintained separate datasets combining clean images with their respective adversarial examples. The ratio between clean and adversarial examples was determined through preliminary experiments to optimize robustness while maintaining clean data performance.

Training Protocol Each model configuration underwent training for 80,000 iterations on an NVIDIA A100 GPU. We implemented early stopping when validation performance showed no improvement over 5,000 consecutive iterations. The learning rate followed a polynomial decay schedule from 0.001 to 0.0001 over the first 40,000 iterations. Performance monitoring was conducted on both clean and adversarial validation sets at 1,000-iteration intervals.

Evaluation Procedure Models were evaluated on three distinct test sets:

- Clean test data to establish baseline performance
- Test data with adversarial examples from their respective training method
- Test data with adversarial examples from all other methods to assess cross-adversarial robustness

RQ2: How effectively does SegRMT enhance segmentation model robustness in self-adversarial testing compared to traditional gradient-based attacks?

Motivation: This research question aims to evaluate which fine-tuned model achieves greater robustness: one fine-tuned using adversarial examples generated by SegRMT, and

Table 5.2 Performance as mIoU(%) of Fine-tuned Models on Adversarial and Clean Datasets

Model	**Clean Dataset**	**Adversarial Testing Datasets**				
		SegRMT	**C&W**	**FGSM**	**PGD10**	**PGD40**
SegRMT	**77.4**%	**53.8**%	68.0%	49.5%	45.0%	46.0%
C&W	76.9%	9.8%	**72.0**%	53.0%	48.6%	51.0%
FGSM	76.8%	10.0%	72.0%	**66.0**%	66.0%	68.0%
PGD10	76.5%	4.0%	72.0%	65.0%	**68.0**%	68.0%
PGD40	76.3%	2.0%	73.0%	62.0%	66.0%	**66.0**%

another fine-tuned using adversarial examples from traditional gradient-based techniques. Specifically, we want to determine if SegRMT's adversarial examples lead to a more robust model when the same attack technique is used to challenge the model post-finetuning.

Approach: We conducted adversarial training on the DeepLabV3 model using adversarial examples generated by various methods, including SegRMT, C&W, FGSM, and PGD. The training dataset was augmented with these adversarial examples alongside the original clean images to challenge the model with a broad spectrum of perturbations. We implemented a systematic training protocol, as described in the Hyperparameter Selection section, which includes information on learning rates, batch sizes, and the proportion of clean to adversarial samples. The model underwent training for more than 80,000 iterations, during which its performance was assessed on both clean and adversarial validation sets. This approach allowed for a thorough evaluation of the model's robustness against the specific adversarial attacks on which it was fine-tuned.

Results: The results presented in Table 5.2 highlight the significant improvements in robustness achieved through adversarial training. The model fine-tuned using C&W adversarial examples exhibited the highest performance increase, with the mIoU improving from 21.7% to 72% on the C&W-generated adversarial dataset. This substantial improvement demonstrates the effectiveness of fine-tuning the model using C&W attacks, significantly enhancing its resilience to this particular form of attack. Similarly, the FGSM fine-tuned model showed a notable increase in performance, with the mIoU rising from 11.3% to 66%. The PGD10 and PGD40 fine-tuned models also demonstrated considerable improvements, with mIoU values increasing from 9.4% to 68% and from 8.5% to 66%, respectively.

The model fine-tuned using adversarial examples created by the SegRMT approach also showed a significant enhancement in performance, with the mIoU improving from 6.4% to

53.8%. While this increase is not as large as that achieved with the C&W fine-tuning, it still demonstrates the overall effectiveness of SegRMT in improving the model's resilience to its own perturbations. These results indicate that adversarial training, regardless of the method used to generate the adversarial examples, substantially improves the robustness of the model. However, the C&W fine-tuned model showed the most substantial improvements, suggesting that this method may offer a particularly effective strategy for adversarial training when the goal is to enhance resistance to specific, well-optimized attacks.

RQ3: How effectively does SegRMT improve segmentation model robustness in cross-adversarial testing versus traditional gradient-based attacks?

Motivation: The aim of this research question is to evaluate the generalization capability of models fine-tuned with SegRMT-generated adversarial examples when tested across a variety of adversarial datasets. Cross-adversarial testing is essential for assessing whether a model's robustness extends beyond the specific attacks it was trained on, thereby demonstrating its ability to defend against a broader range of adversarial perturbations.

Approach: Following the fine-tuning of the DeepLabV3 model using adversarial instances produced by SegRMT, together with conventional gradient-based approaches such as C&W, FGSM, and PGD, we assessed the overall performance of these models on various adversarial datasets. The evaluation aimed to determine whether the robustness gained from training with one type of attack could generalize to other, dissimilar attacks. The specifics of the training process, including hyperparameters and the ratio of clean to adversarial examples, are detailed in the Hyperparameter Selection section. The model's performance was evaluated by measuring its Mean Intersection over Union (mIoU) and Peak Signal-to-Noise Ratio (PSNR), specifically examining its ability to adapt robustly to both gradient-based and non-gradient-based adversarial examples.

Results: When analyzing the general performance of the fine-tuned models across various adversarial datasets, a clear pattern emerges. Models fine-tuned on adversarial examples generated by gradient-based methods, such as FGSM, PGD, and C&W, generally performed well on other gradient-based adversarial datasets. This consistency is likely due to the shared characteristics and similar perturbation patterns among these attacks. However, these same models exhibited poor performance when tested against adversarial examples generated by SegRMT, a metamorphic approach. For instance, models fine-tuned with PGD10 and PGD40

saw their performance on SegRMT-generated adversarial examples degrade significantly, with mIoU dropping from 6.4% in the base model to 4% and 2%, respectively.

Conversely, the model fine-tuned with SegRMT adversarial examples demonstrated better general performance across all types of attacks, including those generated by gradient-based methods. This suggests that the metamorphic testing approach produces more diverse and realistic adversarial examples, enhancing the model's robustness against a broader range of perturbations. The results also indicate that while fine-tuning with specific gradient-based attacks improves resilience against similar perturbations, it may inadvertently reduce robustness against more diverse adversarial examples, such as those generated by SegRMT.

Additionally, the overall performance of the models on clean data experienced only a slight decrease following adversarial training. This minor degradation is a typical outcome in adversarial machine learning, where the trade-off for increased robustness against attacks is a small decrease in accuracy on non-adversarial inputs. The minimal decrease observed suggests that the adversarial training process effectively balanced the need for robustness with the maintenance of performance on clean data.

To determine the statistical significance of the performance disparities between the metamorphic testing tool and the gradient-based adversarial attack methods, we employed the Wilcoxon signed-rank test and calculated Cohen's d effect size.

The Wilcoxon signed-rank test results consistently demonstrate that the tool's fine-tuned model significantly outperforms the gradient-based models on their respective adversarial datasets. The extremely low p-values ($7.27e^{-12}$) obtained for all pairwise comparisons point to highly significant differences, implying that the tool's model exhibits greater robustness across a range of adversarial attack types.

Cohen's d offers a quantitative gauge of the practical significance of the performance differences. The sizable negative Cohen's d value (-4.91) signifies a substantial effect size, wherein the tool's model demonstrates a marked improvement in performance over the gradient-based models. The magnitude of this effect size underscores the practical importance of the performance differences, above and beyond mere statistical significance.

The confluence of Wilcoxon tests and Cohen's d provides compelling evidence that the tool's model is not only statistically superior to gradient-based models in terms of robustness against adversarial attacks but also practically superior, with a large effect size pointing to meaningful performance differences.

The violin plot in Figure 5.3 lends further credence to this analysis by depicting the Intersection over Union (IoU) scores for various images derived from the adversarial examples generated by the gradient-based attack methods. The consistency in the performance of

the model fine-tuned on adversarial examples generated by the metamorphic tool is readily apparent from the plot, as there are no significant outliers.
This finding indicates that the model maintains robust performance across a broad spectrum of adversarial examples, further emphasizing the effectiveness of the metamorphic tool in generating diverse and challenging adversarial examples that enhance the model's robustness.
In sum, the metamorphic tool's capacity to generate varied adversarial examples renders it a valuable asset in preparing models for real-world adversarial scenarios, underscoring the importance of employing diverse adversarial training methods to achieve comprehensive robustness.

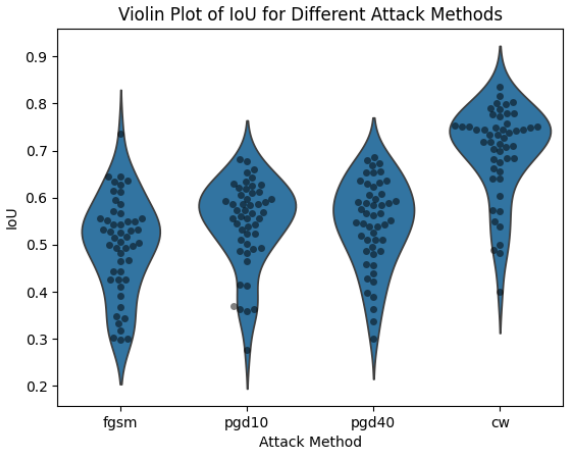


Figure 5.3 Violin Plot of IoU for Different Attack Methods

Overall, the metamorphic tool's ability to generate varied adversarial examples makes it a valuable asset in preparing models for real-world adversarial scenarios, underscoring the importance of employing diverse adversarial training methods to achieve comprehensive robustness.

5.6 Discussion

In RQ1, our findings reveal a significant distinction in the effectiveness of adversarial attacks generated by SegRMT compared to traditional gradient-based methods. Notably, SegRMT produced the lowest mIoU (6.4%) while maintaining a higher PSNR of 24 dB, indicating that the adversarial examples generated by SegRMT are more detrimental to the model's performance. This suggests that SegRMT's adversarial examples, which include a broader range of realistic distortions—both perceptible and imperceptible—are particularly challenging for

the model. The ability of SegRMT to introduce perturbations that retain a high degree of visual integrity while significantly degrading performance underscores its robustness as an adversarial testing approach.

In RQ2, further analysis of Table 2 results for the self-adversarial testing reveals that the C&W attack achieves the highest mIoU (72%) in self-adversarial testing, which correlates with its relatively modest impact on the model's performance, as seen in Table 1. The C&W attack produces the least drop in mIoU , positioning it as the weakest attack among those evaluated. Weaker attacks like C&W tend to generate higher mIoU scores in self-adversarial testing because they produce less impactful adversarial examples. This highlights a critical aspect of evaluating model robustness: weaker adversarial attacks may appear less detrimental in self-adversarial testing, but they also generate less challenging examples, which could lead to overestimated model robustness.

In RQ3, our findings also provide insights into the generalizability of the fine-tuned models. The results in Table 2 under the SegRMT column show that the model fine-tuned on SegRMT adversarial examples consistently outperforms others across all datasets, demonstrating superior robustness to various attacks, including those it has not encountered during training. This consistent performance underscores the robustness and versatility of SegRMT-generated adversarial examples, enabling the model to generalize better and withstand different types of perturbations. In contrast, models fine-tuned on gradient-based attacks, while performing well on adversarial datasets generated by other gradient-based methods, struggle significantly when tested on the SegRMT-generated dataset. This sharp decline in performance highlights a crucial weakness: gradient-based adversarial training enhances robustness against similar perturbations but fails to protect against more diverse, non-gradient-based adversarial examples. In conclusion, SegRMT offers a more comprehensive and robust approach to adversarial testing and training, improving a model's ability to withstand a wide range of unseen adversarial examples and enhancing its overall robustness. On the other hand, while effective against similar attacks, gradient-based adversarial training does not offer the same level of protection against more sophisticated, non-gradient-based adversaries. Therefore, we suggest incorporating SegRMT into the adversarial training process to achieve a more generalized and robust model against a broader spectrum of adversarial threats.

5.7 Threats to validity

As with any research study, it is important to carefully consider and address potential threats to the validity of the findings.
Throughout the design and execution of our study we have taken several steps to mitigate

these threats and ensure the robustness and reliability of our results.

One potential concern is the fairness of the comparison between our metamorphic testing approach and the baseline adversarial attack methods (FGSM, PGD, C&W).

To address this, we have used identical parameter settings ((learning rate, batch size, ratio of clean to adversarial examples) during the adversarial training process for all methods.This ensures that any observed differences in performance can be attributed to the inherent strengths and weaknesses of the approaches themselves, rather than being influenced by external factors.

Another potential issue is the stochastic nature of the genetic algorithm (GA) used in our metamorphic testing approach . To mitigate the influence of randomness on our conclusions, we have employed a rigorous experimental design involving 10 runs with different random seeds and averaging results. This approach helps to make sure that our findings are stable and reproducible and not influenced by chance .

Regarding the generalizability of our finding, we used a widely-accepted and representative dataset and model to conduct our study. The Cityscapes dataset is a standard benchmark for urban scene segmentation, and the DeeplabV3 model with a ResNet-50 backbone represents one of the top-performing models in the domain. While further testing on additional datasets and architectures would certainly be valuable, our choice of experimental materials provides a solid foundation for drawing meaningful conclusions about the effectiveness of our approach.

To facilitate reproducibility and enable other researchers to build upon our work, we have prioritized transparency in reporting our methodology and results. We have provided detailed information about our experimental setup, including hyperparameters and adversarial training procedures, and we have made our code and data publicly available. This allows for independent verification of our findings and promotes the accumulation of knowledge in the field.

Finally, while our approach has been specifically designed and evaluated in the context of robustness assessment for image segmentation models, we believe that it has significant potential for generalization to other domains. The modular structure of our framework and the flexibility of the metamorphic relations used suggest that our approach could be adapted to address similar challenges in tasks such as object detection or medical image analysis. This opens up exciting avenues for future research and highlights the broad impact and applicability of our work.

5.8 Chapter Summary

One of the primary challenges in ensuring the reliability of image segmentation models is their vulnerability to various forms of adversarial attacks and real-world perturbations. While traditional adversarial training approaches like FGSM, PGD, and C&W exist, they often fail to capture the full spectrum of real-world distortions that models may encounter, particularly in safety-critical applications like autonomous driving and medical imaging. To address these limitations, we have introduced a novel framework that combines metamorphic testing with adversarial training, utilizing a genetic algorithm to generate realistic perturbations. Our approach differs from existing methods by focusing on real-world-like distortions rather than purely mathematical adversarial examples. The framework operates by generating test cases through metamorphic relations that simulate realistic scenarios, then incorporating these examples into the training process to enhance model robustness. To evaluate the effectiveness of our approach, we conducted extensive experiments using the DeepLabV3 model on the Cityscapes dataset, comparing our method against established adversarial attack techniques. The results demonstrate that our framework achieves superior performance in improving model robustness, better generalization to unseen adversarial examples, and enhanced reliability across various types of real-world perturbations

CHAPTER 6 CONCLUSION

In this chapter, we conclude the thesis and summarize our findings. In addition, we will discuss the limitations of our studies and the directions for future work.

6.1 Thesis Findings and Conclusions

This thesis addresses the significant issue of resilience in Deep Learning (DL) systems, concentrating on classification and segmentation problems. While DL systems have demonstrated groundbreaking potential across numerous applications, their vulnerability to adversarial perturbations and slight variations in real-world environements increases the risks, especially in safety critical applications. To address these challenges, this thesis presents two important contributions aimed at evaluating and enhancing DL robustness :

1. **Towards Assessing Deep Learning Test Input Generators** The first contribution comprehensively evaluates four state-of-the-art Test Input Generators (TIGs) namely, DeepHunter, DeepFault, AdvGAN, and SinVAD against multiple key dimensions, including fault-revealing capability, naturalness, diversity, and efficiency. Using diverse datasets with varying complexities (MNIST, CIFAR-10, and ImageNet-1K) and pretrained models (LeNet-5, VGG16, and EfficientNetB3), this study identifies trade-offs between TIG performance metrics. The findings shows that generative approaches such as AdvGAN and SinVAD are particularly effective on simpler datasets, while DeepFault demonstrates consistent robustness across varying complexities. However, scalability challenges emerge with high-resolution datasets, highlighting a need for future innovations. These results have a significant impact since they provide a standardized benchmarking framework that offers actionable insights and practical guidelines for selecting appropriate testing tools in classification models. At the same time, the study's limitations—such as scalability issues and its focus exclusively on image classification—indicate that further research is needed to extend these findings to more complex, real-world scenarios and other application domains. To mitigate threats to validity, the study employed default configurations for each TIG. repeated experiments 10 times to account for random variations, and applied statistical significance tests (Wilcoxon test and Vargha–Delaney effect size) to confirm the findings; however, external validity remains limited by the exclusive focus on image classification datasets and models.

2. **Evaluating and Enhancing Segmentation Model Robustness with Metamorphic Testing** While conducting our literature review on computer vision robustness, we observed that while classification tasks had numerous dedicated testing tools, the segmentation domain remained relatively unexplored despite its growing importance in critical applications. This observation, coupled with the increasing deployment of segmentation models in safety-critical scenarios like autonomous driving and medical imaging, motivated us to develop the SegRMT, a novel framework leveraging Metamorphic testing and genetic algorithm to evaluate and inprove the robustness of segmentation models. SegRMT explores the adversarial search space to generate adversarial examples that preserve the visual coherence and effectively expose the model vulnerabilities. Experimental results on the Cityscapes dataset demonstrate SegRMT's superiority over traditional adversarial methods (e.g., FGSM, PGD), both in degrading model performance and enhancing cross-adversarial robustness through adversarial training. The impact of these findings is considerable, as they offer a new, practical tool for improving segmentation model reliability in safety-critical applications. Nevertheless, the limitations of this work include its evaluation on a single dataset and segmentation architecture, as well as the computational overhead introduced by the genetic algorithm, which may challenge broader applicability. Threats to validity were mitigated by using identical parameter settings during adversarial training for all methods, employing a rigorous experimental design with 10 runs using different random seeds, and ensuring reproducibility through transparency in reporting experimental details. Despite these measures, further testing on additional datasets and architectures is necessary to enhance external validity and generalizability.

Together, these contributions present a comprehensive approach for evaluating and enhancing robustness in DL systems. The findings offer actionable insights and practical tools to improve the reliability and trustworthiness of DL models in complex, real-world environments.

6.2 Discussion and Future Work

The main goal of the benchmarking study is to reveal significant trade-offs among TIGs , highlighting strengths and weaknesses. Advgan for example is effective at generating minimal perturbations and maintaining naturalness for simpler datasets but struggles with scalability for high resolution, complex datasets such as imagenet-1k.This limitation suggests a clear direction for future work: enhancing generator and discriminator architectures to better handle complex data distributions. Similarly , Sinvad excels in generating diverse test cases but faces limitations with high resolution images due to the contraints of its VAE architecture often re-

sulting in out of distribution samples. Addressing these constraints through the integration of advanced generative models, such as Vector Quantized VAEs (VQ-VAEs), could improve SinVAD's ability to produce realistic, in-distribution samples for high-resolution datasets. The Perturbation Based Approaches (PBA), such as DeepHunter and DeepFault exhibit complementary strengths : DeepHunter generates more natural looking test cases while DeepFault excels in revealing robustness issues in the tested models. This evaluation underscores the need for TIGs that can balance fault-revealing capability with naturalness across diverse datasets and model complexities. This study also reveals practical challenges in adapting these TIGS to modern architectures and complex datasets. Many TIGs, including those evaluated in this thesis, were originally designed for simple datasets (CIFAR-10, MNIST) and fully connected neural networks (LENET-5, VGG 16). Adapting them to modern architectures and High-resolutions datasets such as EfficientnetB3 and Imagenet-1k required substantial manual modifications. These challenges highlight the need for more modular and scalable TIG frameworks that can easily adapt to the growing complexity of DL systems.
The second contribution, SegRMT expanding the scope of TIGs by targeting segmentation models, demonstrates a balance between adversarial robustness evaluation and enhancement for segmentation models. By leveraging Metamorphic testing and genetic algorithms, SegRMT generates adversarial examples that are both highly effective in exposing the model's vulnerability and visually plausible. However, SegRMT's current application is limited to the segmentations tasks. Testing it's scalability and generalizability to other datasets and architectures represents an intresting avenue for future research. Another important challenge lies in the computational cost associated with SegRMT's genetic algorithm. Optimizing this algorithm or incorporating distributed and parallelized processing techniques could make the framework more scalable and efficient for larger datasets and more complex architectures. Additionally, integrating complementary techniques, such as data augmentation or architectural optimization, could further enhance SegRMT's robustness-enhancing capabilities.

Both contributions highlight the need for a unified robustness framework capable of addressing classification and segmentation tasks. Traditional TIGs excel in generating adversarial scenarios for classification, while SegRMT extends TIG capabilities to segmentation tasks, exposing vulnerabilities and enhancing robustness through adversarial training. A unified framework must retain task-specific strengths, adapt to diverse datasets and architectures, and integrate advanced generative models like VQ-VAEs to bridge methodological gaps.

Additionally, interpretability is crucial. Robustness frameworks should incorporate visualization tools and explainable metrics to make outputs actionable, particularly in safety-critical domains. Testing these frameworks under realistic conditions, accounting for domain shifts and fairness concerns, would further ensure their reliability.

By addressing scalability, modularity, and interpretability, future research can create comprehensive TIG frameworks that evaluate and enhance robustness across DL tasks, enabling safe deployment in real-world environments.

REFERENCES

[1] M. H. M. Noor and A. O. Ige, "A Survey on State-of-the-art Deep Learning Applications and Challenges," 2024. [Online]. Available: https://arxiv.org/abs/2403.17561

[2] S. M. Grigorescu, B. Trasnea, T. T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, pp. 362 – 386, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204744017

[3] T. A. Mim and T. A. Rimi, "A Review on Disease Detection from Medical Images using Machine Learning," *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1437–1441, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249048784

[4] X. Cai, Y. Wen, and J. Liang, "Editorial: Segmentation and classification: theories, algorithms and applications," *Frontiers in Computer Science*, vol. 6, 2024. [Online]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1363578

[5] L. Li, X. Mu, S. Li, and H. Peng, "A Review of Face Recognition Technology," *IEEE Access*, vol. 8, pp. 139 110–139 120, 2020.

[6] J. Villalba-Díez, D. Schmidt, R. Gevers, J. B. O. Meré, M. Buchwitz, and W. Wellbrock, "Deep Learning for Industrial Computer Vision Quality Control in the Printing Industry 4.0," *Sensors (Basel, Switzerland)*, vol. 19, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202711306

[7] A. Mirbeik and N. E. Tavassolian, "Deep Learning for Tumor Margin Identification in Electromagnetic Imaging," in *2023 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (USNC-URSI)*, 2023, pp. 1873–1874.

[8] J. Wang, J. Chen, Y. Sun, X. Ma, D. Wang, J. Sun, and P. Cheng, "RobOT: Robustness-Oriented Testing for Deep Learning Systems," 2021. [Online]. Available: https://arxiv.org/abs/2102.05913

[9] H. Shu and H. Zhu, "Sensitivity Analysis of Deep Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, p. 4943–4950, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1609/aaai.v33i01.33014943

[10] A. Miyajiwala, A. Ladkat, S. Jagadale, and R. Joshi, *On Sensitivity of Deep Learning Based Text Classification Algorithms to Practical Input Perturbations*. Springer International Publishing, 2022, p. 613–626. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-10464-0_42

[11] B. L. Badger, "Depth and Representation in Vision Models," 2023. [Online]. Available: https://arxiv.org/abs/2211.06496

[12] Y. Yu, C. Wang, Q. Fu, R. Kou, F. Huang, B. Yang, T. Yang, and M. Gao, "Techniques and challenges of image segmentation: A review," *Electronics*, vol. 12, no. 5, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/5/1199

[13] B. Emek Soylu, M. S. Guzel, G. E. Bostanci, F. Ekinci, T. Asuroglu, and K. Acici, "Deep-learning-based approaches for semantic segmentation of natural scene images: A review," *Electronics*, vol. 12, no. 12, 2023. [Online]. Available: https://www.mdpi.com/2079-9292/12/12/2730

[14] D. Karunakaran, J. S. Berrio Perez, and S. Worrall, "Generating edge cases for testing autonomous vehicles using real-world data," *Sensors*, vol. 24, no. 1, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/108

[15] V. Riccio and P. Tonella, "When and why test generators for deep learning produce invalid inputs: an empirical study," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1161–1173.

[16] Z. Wang and L. Xu, "Susceptibility of Adversarial Attack on Medical Image Segmentation Models," 2024. [Online]. Available: https://arxiv.org/abs/2401.11224

[17] O. Turnbull and G. Cevora, "Instability of computer vision models is a necessary result of the task itself," 2023. [Online]. Available: https://arxiv.org/abs/2310.17559

[18] J. Kelly, S. A. Zafar, L. Heidemann, J.-V. Zacchi, D. Espinoza, and N. Mata, "Navigating the EU AI Act: A Methodological Approach to Compliance for Safety-critical Products," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, Jun. 2024, p. 979–984. [Online]. Available: http://dx.doi.org/10.1109/CAI59869.2024.00179

[19] Mega, "Replication package," 2025. [Online]. Available: https://mega.nz/folder/nNRBTAaD#QQlGdv3GprV8d2cGg6eNeA

[20] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *ArXiv*, vol. abs/1511.08458, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:9398408

[21] M. Bustreo, C. Beltrán-González, V. Murino, and F. Camozzi, "Recurrent Neural Networks," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:2444500

[22] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *Deep Learning in Science*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:212717965

[23] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, pp. 139 – 144, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1033682

[24] W. Ruan, X. Yi, and X. Huang, "Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications," 2021. [Online]. Available: https://arxiv.org/abs/2108.10451

[25] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.

[26] T. Long, Q. Gao, L. Xu, and Z. Zhou, "A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions," *Computers Security*, vol. 121, p. 102847, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404822002413

[27] C. Li, H. Wang, W. Yao, and T. Jiang, "Adversarial attacks in computer vision: a survey," *J. Membr. Comput.*, vol. 6, no. 2, pp. 130–147, June 2024. [Online]. Available: https://doi.org/10.1007/s41965-024-00142-3

[28] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[29] N. Drenkow, N. Sani, I. Shpitser, and M. Unberath, "A Systematic Review of Robustness in Deep Learning for Computer Vision: Mind the gap?" 2022. [Online]. Available: https://arxiv.org/abs/2112.00639

[30] P. Joshi, M. Z. Shaikh, N. Varshney, and B. Dwivedy, "Robustness Challenges in Deep Learning: Strategies for Enhancing Model Resilience," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, 2024, pp. 757–762.

[31] N. Inkawhich, G. McDonald, and R. Luley, "Adversarial Attacks on Foundational Vision Models," 2023. [Online]. Available: https://arxiv.org/abs/2308.14597

[32] P. Xie, Y. Bie, J. Mao, Y. Song, Y. Wang, H. Chen, and K. Chen, "Chain of Attack: On the Robustness of Vision-Language Models Against Transfer-Based Adversarial Attacks," 2024. [Online]. Available: https://arxiv.org/abs/2411.15720

[33] H. B. Braiek and F. Khomh, "Machine Learning Robustness: A Primer," 2024. [Online]. Available: https://arxiv.org/abs/2404.00897

[34] R. Bouchoucha, H. B. Braiek, F. Khomh, S. Bouzidi, and R. Zaatour, "Robustness assessment of hyperspectral image cnns using metamorphic testing," *Information and Software Technology*, vol. 162, p. 107281, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584923001350

[35] K. Pei, Y. Cao, J. Yang, and S. Jana, "DeepXplore: Automated Whitebox Testing of Deep Learning Systems," in *Proceedings of the 26th Symposium on Operating Systems Principles*, ser. SOSP '17. ACM, Oct. 2017, p. 1–18. [Online]. Available: http://dx.doi.org/10.1145/3132747.3132785

[36] J. Wang, H. Qiu, Y. Rong, H. Ye, Q. Li, Z. Li, and C. Zhang, "Bet: black-box efficient testing for convolutional neural networks," 07 2022, pp. 164–175.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[38] T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, T. H. Tse, and Z. Q. Zhou, "Metamorphic testing: A review of challenges and opportunities," *ACM Comput. Surv.*, vol. 51, no. 1, Jan. 2018. [Online]. Available: https://doi.org/10.1145/3143561

[39] A. Arrieta, "Multi-objective metamorphic follow-up test case selection for deep learning systems," 07 2022, pp. 1327–1335.

[40] P. Patel and A. Thakkar, "The upsurge of deep learning for computer vision applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 538–548, 2020.

[41] A. Paleyes, R.-G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: A survey of case studies," *ACM Comput. Surv.*, vol. 55, no. 6, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3533378

[42] D. "davidad" Dalrymple, J. Skalse, Y. Bengio, S. Russell, M. Tegmark, S. Seshia, S. Omohundro, C. Szegedy, B. Goldhaber, N. Ammann, A. Abate, J. Halpern, C. Barrett, D. Zhao, T. Zhi-Xuan, J. Wing, and J. Tenenbaum, "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," 2024. [Online]. Available: https://arxiv.org/abs/2405.06624

[43] M. Laad, R. Maurya, and N. Saiyed, "Unveiling the Vision: A Comprehensive Review of Computer Vision in AI and ML," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6.

[44] Y. Bi, B. Xue, P. Mesejo, S. Cagnoni, and M. Zhang, "A Survey on Evolutionary Computation for Computer Vision and Image Analysis: Past, Present, and Future Trends," *IEEE Transactions on Evolutionary Computation*, vol. 27, no. 1, p. 5–25, Feb. 2023. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2022.3220747

[45] Z. Chen, C. Wang, and D. J. Crandall, "Semantically Stealthy Adversarial Attacks against Segmentation Models," 2022. [Online]. Available: https://arxiv.org/abs/2104.01732

[46] C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, and S. Zheng, "A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking," 2023. [Online]. Available: https://arxiv.org/abs/2302.14301

[47] C. Kamann and C. Rother, "Benchmarking the Robustness of Semantic Segmentation Models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020, p. 8825–8835. [Online]. Available: http://dx.doi.org/10.1109/CVPR42600.2020.00885

[48] M. K. Ahuja, A. Gotlieb, and H. Spieker, "Testing Deep Learning Models: A First Comparative Study of Multiple Testing Techniques," in *2022 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, Apr. 2022, p. 130–137. [Online]. Available: http://dx.doi.org/10.1109/ICSTW55395.2022.00035

[49] X. Xie, S. See, L. Ma, F. Juefei-Xu, M. Xue, H. Chen, Y. Liu, J. Zhao, B. Li, and J. Yin, "Deephunter: a coverage-guided fuzz testing framework for deep neural networks," 07 2019, pp. 146–157.

[50] H. F. Eniser, S. Gerasimou, and A. Sen, "Deepfault: Fault localization for deep neural networks," in *International Conference on Fundamental Approaches to Software Engineering*. Springer, 2019, pp. 171–191.

[51] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating Adversarial Examples with Adversarial Networks," 2019. [Online]. Available: https://arxiv.org/abs/1801.02610

[52] S. Kang, R. Feldt, and S. Yoo, "SINVAD: Search-based Image Space Navigation for DNN Image Classifier Test Input Generation," 2020. [Online]. Available: https://arxiv.org/abs/2005.09296

[53] Maryam, M. Biagiola, A. Stocco, and V. Riccio, "Benchmarking Generative AI Models for Deep Learning Test Input Generation," 2024. [Online]. Available: https://arxiv.org/abs/2412.17652

[54] V. Riccio and P. Tonella, "When and Why Test Generators for Deep Learning Produce Invalid Inputs: an Empirical Study," 2022. [Online]. Available: https://arxiv.org/abs/2212.11368

[55] Y. Xu, R. Quan, W. Xu, Y. Huang, X. Chen, and F. Liu, "Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches," *Bioengineering*, vol. 11, no. 10, 2024. [Online]. Available: https://www.mdpi.com/2306-5354/11/10/1034

[56] S. Some and V. P. Namboodiri, "Trusting Semantic Segmentation Networks," 2024. [Online]. Available: https://arxiv.org/abs/2406.14201

[57] A. Arnab, O. Miksik, and P. H. S. Torr, "On the Robustness of Semantic Segmentation Models to Adversarial Attacks," 2018. [Online]. Available: https://arxiv.org/abs/1711.09856

[58] L. Halmosi, B. Mohos, and M. Jelasity, "Evaluating the Adversarial Robustness of Semantic Segmentation: Trying Harder Pays Off," 2024. [Online]. Available: https://arxiv.org/abs/2407.09150

[59] J. Gu, H. Zhao, V. Tresp, and P. Torr, "SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness," 2023. [Online]. Available: https://arxiv.org/abs/2207.12391

[60] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, "A Survey on Metamorphic Testing," *IEEE Transactions on Software Engineering*, vol. 42, no. 9, pp. 805–824, 2016.

[61] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th International Conference on Software Engineering*, ser. ICSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 303–314. [Online]. Available: https://doi.org/10.1145/3180155.3180220

[62] T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, T. H. Tse, and Z. Q. Zhou, "Metamorphic Testing," *ACM Computing Surveys (CSUR)*, vol. 51, pp. 1 – 27, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4271578

[63] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, 04 1992. [Online]. Available: https://doi.org/10.7551/mitpress/1090.001.0001

[64] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2890–2896. [Online]. Available: https://aclanthology.org/D18-1316

[65] D. Yang, Z. Yu, H. Yuan, and Y. Cui, "An improved genetic algorithm and its application in neural network adversarial attack," *PLOS ONE*, vol. 17, no. 5, p. e0267970, May 2022. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0267970

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

[67] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[68] H. Geoffrey, D. Li, Y. Dong, E. D. George, and A.-r. Mohamed, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[69] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[70] K. D. Julian, J. Lopez, J. S. Brush, M. P. Owen, and M. J. Kochenderfer, "Policy compression for aircraft collision avoidance systems," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 2016, pp. 1–10.

[71] "Uber's self-driving operator charged over fatal crash," *BBC News*, Sep. 2020. [Online]. Available: https://www.bbc.com/news/technology-54175359

[72] S. . V. b. J. G. Sarlin, Jon, "A false facial recognition match sent this innocent Black man to jail | CNN Business," Apr. 2021. [Online]. Available: https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

[73] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[74] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th international conference on software engineering*, 2018, pp. 303–314.

[75] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *proceedings of the 26th Symposium on Operating Systems Principles*, 2017, pp. 1–18.

[76] X. Xie, L. Ma, F. Juefei-Xu, M. Xue, H. Chen, Y. Liu, J. Zhao, B. Li, J. Yin, and S. See, "Deephunter: a coverage-guided fuzz testing framework for deep neural networks," in *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2019, pp. 146–157.

[77] S. A. A. Shah, M. Beugre, N. Akhtar, M. Bennamoun, and L. Zhang, "Efficient detection of pixel-level adversarial attacks," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 718–722.

[78] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[79] L. Pinheiro Cinelli, M. Araújo Marins, E. A. Barros da Silva, and S. Lima Netto, "Variational autoencoder," in *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer, 2021, pp. 111–149.

[80] S. Li, S. Zhang, G. Chen, D. Wang, P. Feng, J. Wang, A. Liu, X. Yi, and X. Liu, "Towards benchmarking and assessing visual naturalness of physical world adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 324–12 333.

[81] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[82] S. Kang, R. Feldt, and S. Yoo, "Sinvad: Search-based image space navigation for dnn image classifier test input generation," in *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 2020, pp. 521–528.

[83] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[85] N. Krizhevsky, H. Vinod, C. Geoffrey, M. Papadakis, and A. Ventresque, "CIFAR-10 and CIFAR-100 datasets." [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[87] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, 2020.

[88] L. Ma, F. Juefei-Xu, F. Zhang, J. Sun, M. Xue, B. Li, C. Chen, T. Su, L. Li, Y. Liu *et al.*, "Deepgauge: Multi-granularity testing criteria for deep learning systems," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 120–131.

[89] R. Wei, L. Cai, L. Zhao, A. Yu, and D. Meng, "Deephunter: A graph neural network based approach for robust cyber threat hunting," in *Security and Privacy in Communication Networks: 17th EAI International Conference, SecureComm 2021, Virtual Event, September 6–9, 2021, Proceedings, Part I 17*. Springer, 2021, pp. 3–24.

[90] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[91] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[92] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[93] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[94] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[95] B. Luo, Y. Liu, L. Wei, and Q. Xu, "Towards imperceptible and robust adversarial example attacks against neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[96] C. Xie, M. Tan, B. Gong, J. Wang, A. L. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 819–828.

[97] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[98] P. Ralph, N. b. Ali, S. Baltes, D. Bianculli, J. Diaz, Y. Dittrich, N. Ernst, M. Felderer, R. Feldt, A. Filieri *et al.*, "Empirical standards for software engineering research," *arXiv preprint arXiv:2010.03525*, 2020.

[99] J. Wu, M. Zhou, C. Zhu, Y. Liu, M. Harandi, and L. Li, "Performance evaluation of adversarial attacks: Discrepancies and solutions," *arXiv preprint arXiv:2104.11103*, 2021.

[100] A. Shaeiri, R. Nobahari, and M. H. Rohban, "Towards deep learning models resistant to large perturbations," *arXiv preprint arXiv:2003.13370*, 2020.

[101] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.

[102] A. Vargha and H. D. Delaney, "A critique and improvement of the CL common language effect size statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.

[103] M. R. Hess and J. D. Kromrey, "Robust confidence intervals for effect sizes: A comparative study of Cohen'sd and Cliff's delta under non-normality and heterogeneous variances," in *annual meeting of the American Educational Research Association*, vol. 1. Citeseer, 2004.

[104] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[105] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[106] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[107] ——, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[108] B. C. Hu, L. Marsso, K. Czarnecki, R. Salay, H. Shen, and M. Chechik, "If a human can see it, so should your system: Reliability requirements for machine vision components," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 1145–1156.

[109] V. Prabhu, S. Yenamandra, P. Chattopadhyay, and J. Hoffman, "LANCE: Stress-testing Visual Models by Generating Language-guided Counterfactual Images," 2023. [Online]. Available: https://arxiv.org/abs/2305.19164

[110] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, and G. Buttazzo, "On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[111] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, p. 210, 2022.

[112] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," *Expert Systems with Applications*, vol. 198, p. 116815, 2022.

[113] R. Bouchoucha, H. B. Braiek, F. Khomh, S. Bouzidi, and R. Zaatour, "Robustness assessment of hyperspectral image CNNs using metamorphic testing," *Information and Software Technology*, vol. 162, p. 107281, 2023.

[114] X. Wang, Y. Li, C.-J. Hsieh, and T. C. Lee, "Uncovering Distortion Differences: A Study of Adversarial Attacks and Machine Discriminability," *IEEE Access*, 2024.

[115] D. Usynin, D. Rueckert, and G. Kaissis, "Beyond gradients: Exploiting adversarial priors in model inversion attacks," *ACM Transactions on Privacy and Security*, vol. 26, no. 3, pp. 1–30, 2023.

[116] X. Dong, D. Chen, J. Bao, C. Qin, L. Yuan, W. Zhang, N. Yu, and D. Chen, "Greedyfool: Distortion-aware sparse adversarial attack," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 226–11 236, 2020.

[117] X. Xie, Z. Zhang, T. Y. Chen, Y. Liu, P.-L. Poon, and B. Xu, "METTLE: A metamorphic testing approach to assessing and validating unsupervised machine learning systems," *IEEE Transactions on Reliability*, vol. 69, no. 4, pp. 1293–1322, 2020.

[118] S. H. Santos, B. N. C. Da Silveira, S. A. Andrade, M. Delamaro, and S. R. Souza, "An experimental study on applying metamorphic testing in machine learning applications," in *Proceedings of the 5th Brazilian Symposium on Systematic and Automated Software Testing*, 2020, pp. 98–106.

[119] X. Xie, J. W. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," *Journal of Systems and Software*, vol. 84, no. 4, pp. 544–558, 2011.

[120] Y. Ma, Y. Pan, and Y. Fan, "Metamorphic Testing for the Medical Image Classification Model," in *2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. IEEE, 2022, pp. 340–346.

[121] Y. Pan, H. Ao, and Y. Fan, "Metamorphic testing for autonomous driving systems in fog based on quantitative measurement," in *2021 IEEE 21st International Conference*

on Software Quality, Reliability and Security Companion (QRS-C). IEEE, 2021, pp. 30–37.

[122] Y. Sheng, J. Yang, Y. Lin, W. Jiang, and L. Yang, "Toward Fair Ultrasound Computing Tomography: Challenges, Solutions and Outlook," in *Proceedings of the Great Lakes Symposium on VLSI 2024*, 2024, pp. 748–753.

[123] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[124] L. Applis, A. Panichella, and R. Marang, "Searching for quality: Genetic algorithms and metamorphic testing for software engineering ml," in *Proceedings of the Genetic and Evolutionary Computation Conference*, ser. GECCO '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1490–1498. [Online]. Available: https://doi.org/10.1145/3583131.3590379

[125] J. Ayerdi, V. Terragni, G. Jahangirova, A. Arrieta, and P. Tonella, "GenMorph: Automatically Generating Metamorphic Relations via Genetic Programming," *IEEE Transactions on Software Engineering*, vol. 50, no. 7, pp. 1888–1900, 2024.

[126] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[127] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[128] V. Knauthe, A. Rak, T. Wirth, T. Pöllabauer, S. Metzler, A. Kuijper, and D. W. Fellner, "Transparency Distortion Robustness for SOTA Image Segmentation Tasks," 2024. [Online]. Available: https://arxiv.org/abs/2405.12864

[129] V. Gorade, S. Mittal, D. Jha, R. Singhal, and U. Bagci, "Harmonized spatial andnbsp;spectral learning fornbsp;generalized medical image segmentation," in *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XIII*. Berlin, Heidelberg: Springer-Verlag, 2024, p. 178–193. [Online]. Available: https://doi.org/10.1007/978-3-031-78201-5_12

APPENDIX A CO-AUTHORSHIP

Earlier studies in the thesis were published/submitted as follows:

- **Seif Mzoughi***, Ahmed Haj Yahmed*, Mohamed Elshafei, Foutse khomh, Diego Elias Costa, Towards Assessing Deep Learning Test Input Generators, *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering (EASE)*, 2025
- **Seif Mzoughi***, Mohamed Elshafei, Foutse Khomh, Evaluating and Enhancing Segmentation Model Robustness with Metamorphic Testing, in *Journal of Systems and Software (JSS)*, 2024