Pre-trained Model Guided Fine-Tuning for Zero-Shot Adversarial Robustness

|  |  |
| --- | --- |
| John Joseph Zagariah Daniel  (300382174)  Student , Dept. of EECS,  University of Ottawa  jzaga079@uottawa.ca | Pratish Pushparaj  (300375330)  Student , Dept. of EECS,  University of Ottawa  ppush035@uottawa.ca |

*Abstract*—The increasing reliance on vision-language models like CLIP for real-world applications, such as autonomous systems, raises concerns about their susceptibility to adversarial perturbations. While the original PMG-AFT methodology effectively improves adversarial robustness for classification tasks, its applicability to image-to-text description tasks remains unexplored. In this work, we adapt PMG-AFT to address the challenge of generating robust textual descriptions of attributes such as vehicle types, weather, and traffic conditions from scene images, using datasets like KITTI and BDD10k. By framing the description task as a classification problem—where an image is matched to its corresponding text—we analyze the role of robustness loss, generalization loss, and regularization loss in balancing clean accuracy and adversarial robustness. We further question the generalizability of the PMG-AFT loss function across other tasks and domains. Specifically, we modify the regularization loss to use cosine similarity, improving its flexibility for aligning adversarial and clean embeddings. Through ablation studies, we highlight the impact of each loss term and demonstrate that the modified loss function achieves 61.33% adversarial accuracy while preserving clean accuracy at 28.07%. This raises a broader question: *Can language-driven robustness techniques be extended beyond image classification to more complex downstream tasks?* Our findings provide insights into fine-tuning pre-trained vision-language models for robustness in diverse applications, paving the way for exploring similar strategies in other domains.

Keywords—Vision-Language-Models, autonomous driving, Adversarial robustness, Image-to-Text Description,

# Introduction

In recent years, vision-language models have revolutionized the field of computer vision by enabling a wide range of tasks that combine image understanding with natural language processing. Models such as CLIP (Contrastive Language-Image Pretraining) [1] have demonstrated remarkable zero-shot performance across tasks like image classification and scene understanding, achieving impressive generalization without task-specific fine-tuning. CLIP aligns images and textual descriptions into a shared embedding space, making it suitable for downstream applications like image-to-text description. Such tasks are particularly critical for autonomous driving systems, where images must be described in terms of specific attributes, including vehicle types, traffic conditions, and weather.

However, despite their success, vision-language models remain vulnerable to adversarial perturbations—small, imperceptible changes to input images that can drastically degrade model performance. These vulnerabilities pose significant risks for real-world applications, such as autonomous vehicles, where robust performance is essential to ensure safety and reliability. Addressing this issue requires fine-tuning methods that enhance adversarial robustness without compromising the model's generalization ability to clean (unperturbed) inputs.

The original Pre-trained Model Guided Adversarial Fine-Tuning (PMG-AFT) framework [2] was proposed to improve adversarial robustness for image classification tasks. PMG-AFT introduces a dual-branch fine-tuning strategy that balances: (1) Robustness Loss (*Lrobust)*: Ensuring correct classification of adversarial examples. (2) Generalization Loss (*Lgeneral*): Aligning clean and adversarial predictions to retain generalization. (3) Regularization Loss (*Lclean*): Aligning adversarial features with clean features to prevent overfitting.

While PMG-AFT achieves strong results in **image classification**, its applicability to more complex **image-to-text description** tasks remains unexplored. Image-to-text description involves generating textual descriptions for images, focusing on attributes such as **vehicle types**, **traffic density**, and **weather conditions**. In this work, we frame the description task as a **classification problem**, where the model selects the most relevant textual description from a set of candidate descriptions based on the input image. Extending PMG-AFT to this domain presents unique challenges, as fine-tuning must account for both adversarial robustness and the alignment of visual and textual representations.

In this paper, we adapt PMG-AFT for **robust image-to-text description** using datasets such as **KITTI** and **BDD10k**, which are widely used in the autonomous driving domain. To ensure high-quality textual labels, we generate scene descriptions using **BLIP** [3] and perform **extensive manual verification**. Additionally, we modify the regularization loss (*Lclean*) to use **cosine similarity** instead of mean squared error (MSE). This modification aims to provide a more flexible alignment between adversarial and clean embeddings, improving robustness without overly constraining the model.

Through a series of experiments and ablation studies, we demonstrate the following key findings: (1) The **robustness loss** (*Lrobust)* and **generalization loss** (*Lgeneral*) are critical for achieving adversarial robustness. (2) Ablation studies highlight the relative contributions of each loss component to the overall performance.

This work raises important questions about the generalizability of adversarial robustness techniques across different tasks. Specifically, we ask: *Can strategies like PMG-AFT, originally designed for classification, be effectively extended to other vision-language tasks such as image-to-text description?* Our results provide initial insights into this question, highlighting the potential for language-driven robustness techniques to address complex downstream applications.

The rest of the paper is organized as follows. Section 2 presents the methodology, including the PMG-AFT framework and our modifications. Section 3 details the experimental setup, datasets, and evaluation metrics. Section 4 discusses results and ablation studies, and Section 5 concludes the paper with key findings and future directions.

# Methodology

This section presents the approach for improving adversarial robustness in image-to-text description tasks using the PMG-AFT framework. The key components include the problem formulation, the PMG-AFT architecture, and the process for generating adversarial examples.

## Problem Formulation

The image-to-text description task is framed as a classification problem, where the goal is to match an input image 𝑥 to its most relevant textual description from a predefined set of candidate descriptions . The predicted label ​, corresponding to the text description with the highest similarity score. The vision-language model encodes the image and text descriptions into a shared embedding space. The similarity between the image embedding *ν* and each text embedding *ti*​ is computed as the dot product.

Similarity(x,ti) = ImageEncoder(x) ‧ TextEncoder(ti)

The final prediction ​, given by:

​ = argmax(Similarity(x,ti))

## PMG-AFT Framework

The PMG-AFT framework fine-tunes a pre-trained vision-language model (CLIP) to improve adversarial robustness without compromising generalization. The methodology involves two key components: **Robustness Branch** and **Generalization Branch**, along with a **Regularization Loss**. The overall pipeline is illustrated in **Figure 1**.

**Robustness Branch:** The robustness branch focuses on ensuring correct classification of adversarial examples.

**1. Adversarial Image**: Perturbed versions of the clean input image xadv​ are generated using the **PGD attack** (described in Section 2.3).

**2. Trainable Encoder**: The adversarial image xadv​​ is passed through the **trainable target image encoder** to produce adversarial features.

**3. Robustness Loss**: The similarity between adversarial features and text embeddings is computed, and the **cross-entropy loss** is minimized to align adversarial outputs with the ground truth labels Y:

*Lrobust* = CrossEntropy(Similarity(xadv, t) ,y)

**Generalization Branch:** The generalization branch ensures alignment between adversarial and clean predictions to maintain the model’s generalization capability.

**1. Frozen Encoder**: The adversarial image xadv​​ is passed through a **frozen original image encoder** to produce the original adversarial features.

**2. KL Divergence**: The adversarial logits from the **trainable encoder** are aligned with the logits from the **frozen encoder** using **KL Divergence**:

*Lgeneral* = DKL(softmax(logitsadv),softmax(logitsoriginal)).

**Regularization Loss:** The regularization loss ensures similarity between adversarial and clean features, preventing overfitting. In the primary implementation, **Mean Squared Error (MSE)** is used to align the logits:

*Lclean* = MSE(logitsadv, logitsclean).

As part of the ablation study, we explored replacing MSE loss with cosine similarity:

Lcleancosine  = 1 – CosineSimilarity(logitsadv, logitsclean).

**Total Loss:**

The final loss function combines all three components with weights α and β:

*L=Lrobust+αLgeneral+βLclean*

## Adversarial Example Generation

To train the robustness branch, adversarial images are generated using the **Projected Gradient Descent (PGD)** attack. PGD iteratively perturbs the clean input image x to maximize the loss while ensuring the perturbation remains bounded. The process is as follows:

**1. Initialization:** The adversarial image is initialized as the clean input: xadv(0) = x.

**2. Iterative Perturbation:** At each step k, the adversarial image is updated using the gradient of the loss function L:

xadv(k+1) = Projℬ(x,ℰ) ( xadv(k) + 𝛂‧sign(∇xadv(k) L) )

where:

𝛂 – Step Size; ℰ - Perturbation Budget;

Projℬ(x,ℰ)  - ensures the adversarial image remains within the allowable ℰ -ball.

**3. Stopping Condition:** The perturbation process stops after K iterations, and the resulting adversarial image xadv​ is used to compute the losses.

**A diagram of a software process

Description automatically generated with medium confidence**

Figure 1 illustrates the entire pipeline, showing the flow of clean and adversarial images, the frozen and trainable image encoders, text embeddings, and the computation of *Lrobust* ​, *Lgeneral* ​, and *Lclean*

# EXPERIMENT

This section describes the experimental setup, including the datasets, preprocessing, training configuration, evaluation metrics, and ablation study design. The experiments are conducted to validate the proposed **PMG-AFT** framework for improving adversarial robustness in the image-to-text description task.

## Dataset and Label Generation

The experiments are conducted on a combined dataset created using the **KITTI** and **BDD10K** datasets. Both datasets are widely used in the field of autonomous driving research and contain real-world images that reflect urban and natural driving scenarios. KITTI provides high-quality images depicting self-driving car scenes, including vehicles, pedestrians, and road conditions in urban settings. In contrast, BDD10K introduces a broader range of diversity, covering weather conditions such as clear, rainy, and foggy scenes, as well as variations in time of day, including daytime and nighttime images. This combination provides a robust benchmark for evaluating both clean and adversarial performance across varied environmental conditions.

To generate textual descriptions for the images, we utilize **BLIP**, a pre-trained vision-language model capable of producing natural language captions for images. These descriptions focus on key scene attributes such as vehicle types (e.g., cars, buses, and trucks), traffic conditions, pedestrian activity, and weather. Extensive manual verification is performed on the generated descriptions to ensure accuracy and quality, as noisy or irrelevant labels could adversely impact the performance of the proposed model. After combining the two datasets, we construct a single dataset containing **4500 images** for training and **1500 images** for testing. The unified dataset provides a balanced representation of scene attributes while ensuring sufficient data for evaluating adversarial robustness.

Prior to training, all images are preprocessed to match the input format required by the **CLIP ViT-B/32** model. Each image is resized to a fixed resolution of 224×224 pixels. Furthermore, images are normalized using CLIP-specific mean [0.481,0.457,0.408] and standard deviation [0.268,0.261,0.275] values to align them with the distribution of the pre-trained model's training data. These preprocessing steps ensure compatibility with the model and facilitate effective fine-tuning.

## Training Configuration

The model is trained using the combined dataset with adversarial examples generated through the **Projected Gradient Descent (PGD)** attack. The adversarial examples are crafted by iteratively perturbing clean input images within a bounded ϵ -ball to maximize the loss. The perturbation budget ϵ is set to 0.03, with a step size of 0.01 and a total of 10 iterations. The training is performed using the **Adam optimizer** with a fixed learning rate of 1×10−4. The batch size is set to 4, and the model is trained for 5 epochs to ensure convergence. The generalization loss is weighted by α=0.7, while the regularization loss is weighted by β=0.3. These weights are selected based on preliminary experiments to balance the trade-off between robustness and generalization.

## Training Configuration

The performance of the proposed framework is evaluated using two primary metrics: **clean accuracy** and **adversarial accuracy**. Clean accuracy measures the model's performance on unperturbed input images, while adversarial accuracy evaluates its robustness to adversarial examples generated using the PGD attack. The classification task involves predicting the most relevant textual description for a given image from a set of candidate descriptions. These metrics provide a comprehensive assessment of the model’s ability to align image and text features under clean and perturbed conditions.

In addition to evaluating the overall performance, an ablation study is conducted to analyze the relative contributions of the three loss components. Specifically, the experiments include removing the robustness loss (*Lrobust​),* the generalization loss *(Lgeneral​),* and the regularization loss *(Lclean​)* individually to observe their impact on clean and adversarial accuracy. Furthermore, an additional experiment explores replacing the MSE-based regularization loss with cosine similarity to evaluate its effect on feature alignment and adversarial robustness. These studies provide insights into the significance of each loss term in the proposed framework.

# RESULTS AND DISCUSSION

This section presents the results of the experiments conducted to evaluate the proposed adaptation of the PMG-AFT framework for adversarial robust image-to-text description tasks. The results are interpreted in the context of the problem being addressed, highlighting the contributions of the robustness loss, generalization loss, and regularization loss. Additionally, the impact of hyperparameter tuning and substituting MSE loss with cosine similarity is discussed.

To understand the effect of loss weighting on adversarial robustness, experiments were conducted with different values of α (generalization loss weight) and β (regularization loss weight). The results are summarized in **Table 1**.

|  |  |  |
| --- | --- | --- |
| Configuration | Clean Accuracy | Adversarial Accuracy |
| α=0.7,β=0.3 | 0.2807 | 0.6113 |
| α=1.0,β=1.0 | 0.2807 | 0.4567 |
| α=0.1,β=0.1 | 0.2807 | 0.4567 |
| α=0.3,β=0.7 | 0.2807 | 0.4567 |

Table 1: Effect of Loss Weighting on Clean and Adversarial Accuracy

The configuration with α=0.7 and β=0.3 achieves the highest adversarial accuracy of **61.13%** while maintaining clean accuracy at **28.07%**. Increasing both weights to α=1.0 and β=1.0 significantly reduces adversarial accuracy to **45.67%**, suggesting that excessive emphasis on generalization and regularization interferes with the model's ability to distinguish adversarial features. Similarly, when the loss weights are reduced to 0.1 for both terms, the adversarial accuracy remains low at **45.67%**, indicating that insufficient weighting of these terms fails to align adversarial and clean logits effectively. These results highlight the importance of carefully tuning α and β to achieve an optimal trade-off.

To analyze the contribution of each loss component, experiments were performed by removing one loss term at a time while keeping the others intact. The results are summarized in **Table 2**.

The results demonstrate that removing the **robustness loss** *(Lrobust​)* leads to a significant drop in adversarial accuracy to **45.67%**, confirming its critical role in enabling the model to correctly classify adversarial examples. Similarly, the absence of the **generalization loss** *(Lgeneral​)* also reduces adversarial accuracy to **45.67%**, indicating that aligning adversarial logits with clean logits is essential for retaining the model's generalization ability.

|  |  |  |
| --- | --- | --- |
| Configuration | Clean Acc | Adversarial Acc |
| Baseline | 0.2807 | 0.6113 |
| Without Lclean​ | 0.2807 | 0.6273 |
| Without Lrobust​ | 0.2807 | 0.4567 |
| Without Lgeneral​ | 0.2807 | 0.4567 |

Table 2: Ablation Study Results

Interestingly, removing the **regularization loss** *(Lclean​)* results in a slight improvement in adversarial accuracy to **62.73%**, suggesting that the regularization term may impose overly restrictive constraints on the alignment of clean and adversarial features. This observation indicates that the robustness and generalization branches play dominant roles in achieving adversarial robustness, while regularization contributes marginally in this context.

Given the slight improvement observed when removing *Lclean*, we hypothesized that the **MSE loss** might overly constrain the alignment of adversarial and clean logits. This led us to explore **cosine similarity** as an alternative for regularization. Unlike MSE, which enforces magnitude-based constraints, cosine similarity focuses on the angular relationship between vectors. The assumption was that cosine similarity would provide a more relaxed alignment, allowing the model to adapt better to adversarial perturbations. The results show that substituting MSE loss with cosine similarity leads to a negligible improvement in adversarial accuracy to 61.33%, while clean accuracy remains unchanged at 28.07%. This indicates that the MSE-based regularization is already sufficient for aligning adversarial and clean features in this framework. Furthermore, the pre-trained CLIP model produces normalized embeddings, which naturally exhibit cosine-like properties. As a result, explicitly enforcing cosine similarity provides minimal additional benefit. These findings suggest that while cosine similarity offers a conceptually more flexible alignment, its practical impact in this setting is limited. MSE loss remains a strong baseline for regularization within the PMG-AFT framework.

The experimental results demonstrate that the robustness loss *(Lrobust)* and generalization loss *(Lgeneral)* are critical for improving adversarial robustness, with their removal resulting in a significant drop in adversarial accuracy. The regularization loss *(Lclean)* plays a comparatively minor role,

as its removal yields a slight improvement, suggesting it may impose overly restrictive constraints.

To relax these constraints, we explored replacing MSE loss with cosine similarity, hypothesizing that angular alignment would offer greater flexibility. However, the results showed negligible improvement, with adversarial accuracy increasing marginally to 61.33%. This indicates that MSE remains a strong choice for regularization, given the normalized nature of CLIP embeddings.

The clean accuracy remained constant at 28.07% across all experiments, highlighting a trade-off between robustness and performance on clean inputs. These findings emphasize the importance of loss formulation and weighting for achieving robust performance while raising questions about further generalization across tasks and domains.

# CONCLUSION AND FUTURE DIRECTIONS

In this paper, we adapted the **PMG-AFT framework** to the image-to-text description task, focusing on generating robust textual descriptions of scene attributes such as vehicle types, weather, and traffic conditions. By framing the description task as a classification problem, we demonstrated the roles of robustness loss, generalization loss, and regularization loss in improving adversarial robustness. Our results achieved an adversarial accuracy of **61.33%** while maintaining a clean accuracy of **28.07%**, with ablation studies validating the importance of the robustness and generalization losses. While the MSE-based regularization loss proved effective, the use of cosine similarity provided limited benefits, indicating the need for further exploration of flexible alignment strategies. The clean-adversarial performance trade-off highlights a critical challenge in the current formulation.

Future work will focus on dynamic loss weighting techniques to balance clean and adversarial accuracy, extending the framework to downstream tasks such as image captioning and object detection, and evaluating robustness on broader datasets. These efforts will further generalize the PMG-AFT methodology, enhancing its applicability to real-world vision-language systems.

##### References

1. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.
2. S. Wang, J. Zhang, Z. Yuan, and S. Shan, “Pre-trained Model Guided Fine-Tuning for Zero-Shot Adversarial Robustness,” CVPR, 2024
3. J. Li et al., "BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding and Generation," NeurIPS, 2022.
4. A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? The KITTI vision benchmark suite,” CVPR, 2012
5. F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “BDD100K: A diverse driving dataset for heterogeneous multitask learning,”