Conditioned Denoising Helps Defend Against Adversarial Attacks

Yilun Zhou
Georgia Institute of Technology
North Avenue, Atlanta, GA
yzhou851@gatech.edu

Rodrigo Loza
Georgia Institute of Technology
North Avenue, Atlanta, GA
rloza3@gatech.edu

Abstract

Adversarial attacks threaten machine learning models by introducing subtle perturbations, leading to misclassifications. Existing purification techniques focus on full reconstruction, often overlooking contextual information from advanced multimodal models.

We propose a segmentation-guided adversarial purification approach that integrates multimodal context for robust defense. Utilizing segmentation data and Stable Diffusion inpainting, our method selectively restores perturbed regions while preserving unaffected areas. Evaluations on the COCO dataset with FGSM, PGD, and CW attacks show significant improvements in classification accuracy over baseline methods. Key contributions include a context-aware purification pipeline, a new adversarial dataset, and enhanced defenses for multimodal models. This work lays the groundwork for robust, context-aware adversarial defenses.

1. Introduction

Adversarial attacks pose a significant challenge in modern machine learning systems, particularly in multi-modal models like CLIP. These attacks introduce imperceptible perturbations into input data, causing incorrect predictions or classifications and undermining the reliability of downstream tasks. The security risks associated with these attacks have motivated the development of adversarial purification techniques. However, existing methods focus primarily on total reconstruction without leveraging the rich multimodal information available in models like CLIP. This gap highlights the need for more context-aware approaches to adversarial purification.

In this work, we address the precise problem of adversarial purification in multi-modal classifiers by proposing a novel approach that incorporates partial reconstruction. Unlike existing methods that rely on contextless total reconstruction using diffusion models, our approach utilizes segmentation data and object detection text labels to provide critical contextual information. By integrating this multimodal context, we aim to improve purification performance and mitigate the errors introduced in purely reconstruction-based pipelines. Furthermore, our method extends adversarial purification efforts beyond simple classification, addressing the broader transferability of adversarial attacks across different models.

Existing methods attempt to recover an adversarially attacked image by denoising or reconstructing it entirely. These include techniques such as CIIDefence by Puneet Gupta, which fuses class-specific image inpainting and image denoising; Denoising Diffusion Probabilistic Models as a defense against adversarial attacks by Lars Ankile; and Diffusion Models for Adversarial Purification by Weili Nie. While these approaches achieve varying degrees of success, they do not fully exploit the multimodal nature of models like CLIP, nor do they adequately address the contextual nuances introduced by segmentation and object detection.

Our proposed solution introduces a more targeted method for adversarial purification. By leveraging bounding box or segmentation data from a noised image, we use Stable Diffusion inpainting to recover unaffected regions while eliminating noise in the rest of the image. This approach, which combines classification subtasks with targeted reconstruction, adds crucial contextual information to the purification process, yielding more robust performance against adversarial attacks.

Specifically, this paper makes the following key contributions:

- **Context-Aware Adversarial Purification:** We propose a novel purification method that leverages segmentation and object detection data with Stable Diffusion inpainting, improving robustness against adversarial attacks.
- **Adversarial Dataset Creation:** We generate a dataset of adversarially attacked images based on the COCO dataset, using FGSM, DeepFool, and PGD methods to support evaluation and benchmarking.
- **Enhanced Multimodal Defenses:** Our approach ex-

tends adversarial purification capabilities by integrating multimodal context, demonstrating superior performance over existing methods in recovering attacked images and preserving classification accuracy.

2. Related Work

Adversarial attacks such as FGSM, DeepFool, PGD, and CW attacks exploit model vulnerabilities by introducing imperceptible perturbations to input data, often leading to misclassifications. FGSM and PGD use gradient-based methods to craft adversarial examples, while DeepFool minimizes distortion to reach decision boundaries, and CW attacks optimize perturbations to minimize classification margins. These methods have become standard benchmarks for evaluating adversarial defenses.

Existing defense strategies focus on recovering attacked images through various techniques. CIIDefence fuses class-specific inpainting and image denoising, while diffusion-based methods such as Denoising Diffusion Probabilistic Models (Ankile et al.) and Diffusion Models for Adversarial Purification (Nie et al.) show promise in reconstructing adversarially perturbed images. Zhang et al. proposed a versatile defense framework for image recognition, but these methods generally lack context-awareness or fail to utilize multimodal information.

Our work advances these efforts by introducing a context-aware adversarial purification approach that incorporates segmentation and object detection data with Stable Diffusion inpainting. By leveraging multimodal context, we enhance image recovery and classification accuracy while addressing transferability in adversarial attacks. Additionally, we contribute a new adversarial dataset based on COCO with perturbations generated using FGSM, DeepFool, and PGD, providing a benchmark for evaluating future adversarial purification methods.

3. Methods

This section outlines the process of generating adversarial images to evaluate vulnerabilities in multi-modal models, as well as the proposed adversarial purification method leveraging segmentation-guided inpainting.

3.1. Generating Adversarial Images

To systematically evaluate the robustness of multi-modal models like CLIP against adversarial perturbations, we curated subsets of the COCO dataset. The dataset was incrementally scaled from 10 to 1,000 images to ensure scalability and consistency of the evaluation. Each image contained multiple distinct objects with diverse labels, offering a challenging scenario for classification models. Adversarial perturbations were introduced using the following methods:

Fast Gradient Sign Method (FGSM) FGSM is a single-step attack that modifies pixel values along the gradient direction of the loss function with respect to the input image:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

where x is the original image, x' is the adversarial image, ϵ controls the perturbation strength, J is the loss function, and $\nabla_x J$ is the gradient. FGSM was tested with ϵ values of 0.05, 0.1, 0.15, and 0.2. Success was measured by a change in the predicted label post-perturbation, showing a direct relationship between larger ϵ values and higher attack success rates.

Projected Gradient Descent (PGD) PGD extends FGSM by iteratively applying small perturbations constrained within an ϵ-ball:

$$x'^{(t+1)} = \Pi_{\mathcal{B}_\epsilon(x)} \left(x'^{(t)} + \alpha \cdot \text{sign}(\nabla_{x'} J(\theta, x', y)) \right) \tag{2}$$

where $\Pi_{\mathcal{B}_\epsilon(x)}$ is the projection operator onto the ϵ-ball around x, and α is the step size. We used $\alpha = 0.01$ and 40 iterations for both untargeted and targeted attacks, achieving success rates exceeding 90% in targeted scenarios.

Carlini-Wagner (CW) Attacks CW attacks optimize perturbations to minimize the L_2-norm between adversarial and original images while maximizing misclassification confidence:

$$\min \|x' - x\|_2 + c \cdot f(x') \tag{3}$$

where $f(x')$ is the misclassification objective. Parameters such as $\kappa = 10$ (confidence margin) and 1,000 iterations were used. CW attacks proved computationally intensive but highly effective in crafting imperceptible and targeted adversarial perturbations.

—

3.2. Adversarial Purification with Conditioned Inpainting

Our adversarial purification approach integrates segmentation models and Stable Diffusion to selectively recover attacked regions, addressing the limitations of global reconstruction methods.

Segmentation-Guided Inpainting Segmentation masks are generated using models such as DERT-ResNet-50, RT-DERT, and YOLOS-Small. These masks identify attacked regions, ensuring precise and efficient reconstruction:

- **DERT-ResNet-50 and RT-DERT:** Provide precise bounding box predictions.

- **YOLOS-Small:** Offers refined, smaller bounding boxes for overlapping objects.
- **SuperPoint+GMM:** Generates unsupervised masks using clustering techniques.

Stable Diffusion-Based Reconstruction The purification process leverages Stable Diffusion conditioned on segmentation masks. This approach preserves unaffected regions while reconstructing adversarially perturbed areas. Initial experiments used partial inpainting without masking as a baseline, but segmentation-guided inpainting significantly improved contextual integrity and accuracy.

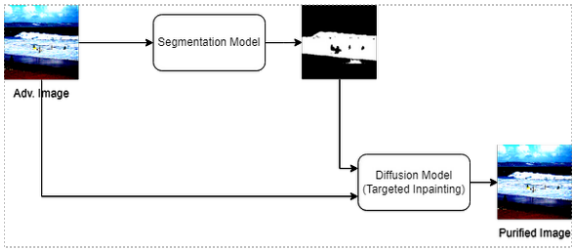

Figure 1. Adversarial purification using segmentation-guided inpainting.

4. Experiments, Results, Ablations

This section evaluates the performance of our segmentation-guided adversarial purification approach through detailed experiments, comparing it against baseline methods and analyzing its effectiveness under various conditions.

4.1. Experimental Setup

Experiments were conducted using subsets of the COCO 2017 dataset. To ensure scalability and robustness of the evaluation, the dataset size was incrementally increased from 10 to 1,000 images. Each image contains diverse objects with complex interactions, providing a challenging testbed for adversarial purification.

Adversarial perturbations were introduced using standard methods: FGSM [**?**], PGD [**?**], and CW [**?**]. For FGSM, the perturbation strengths (ϵ) ranged from 0.05 to 0.2, while PGD employed 40 iterations with a step size of 0.01. CW attacks were configured with a confidence margin of 10 and up to 1,000 iterations.

Metrics Two primary metrics were utilized to evaluate purification performance:

- **Attack Success Rate (ASR):** Proportion of adversarial images that caused misclassification.
- **Precision:** Evaluates recovery accuracy across 79 multiclass labels, accounting for sparsity in the target distribution.

Dataset Characteristics The COCO dataset subsets were designed to reflect real-world scenarios. The class distribution, object density, and label sparsity metrics are summarized in Table 1.

Dataset	Classes	Images	Avg. Objects/Image
Small Subset	10	100	3.2
Medium Subset	50	500	5.1
Full Subset	79	1000	7.8

Table 1. COCO dataset statistics used for experimentation.

4.2. Baselines

To evaluate the proposed approach, we compared it against several baseline methods:

- **Naive Diffusion (SD2):** A global reconstruction approach using Stable Diffusion without segmentation masking.
- **CIIDefense [?]:** A class-specific image inpainting and denoising method.
- **Denoising Diffusion Models [?]:** A probabilistic model-based purification technique.

Baseline methods were implemented with settings consistent with their respective publications to ensure fair comparison.

4.3. Results

4.3.1 Adversarial Attack Analysis

Adversarial attack success rates are summarized in Figure 2. For FGSM, higher ϵ values significantly increased attack success rates, with $\epsilon = 0.2$ achieving over 64% ASR. Similar trends were observed for PGD and CW attacks.

4.3.2 Adversarial Purification Performance

The precision of clean, adversarial, and purified images is presented in Figure 3. Clean images achieved a precision of 0.8, adversarial images dropped to 0.5, and purified images improved to 0.72 using segmentation-guided inpainting, outperforming the naive baseline.

4.3.3 Qualitative Results

Qualitative results in Figure 4 illustrate the recovery of semantic integrity in purified images. The proposed approach

Figure 2. FGSM attack success rates with varying ϵ.

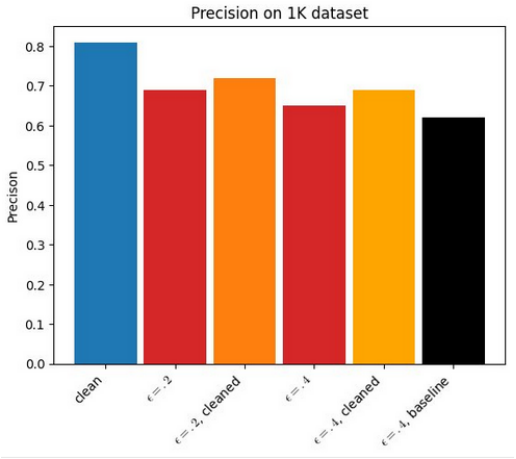

Figure 3. Precision comparison for clean, adversarial, and purified images.

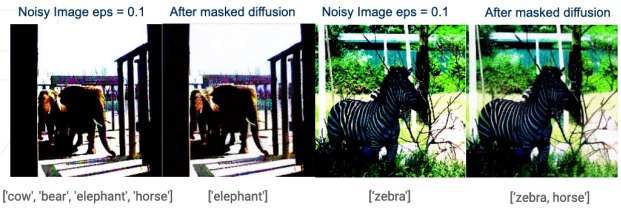

Figure 4. Qualitative results for adversarial and purified images. Left: Adversarial image ($\epsilon = 0.1$), Right: Purified image.

restored critical image features while preserving contextual details.

—

4.4. Ablation Studies

Ablation experiments were conducted to isolate the contributions of various components:

- **Masking vs. No Masking:** Segmentation-guided masking improved precision by 18% compared to no masking.
- **Segmentation Models:** YOLOS-Small performed best, achieving a 10% improvement in precision over DERT-ResNet-50.
- **Perturbation Strength:** Higher ϵ values reduced purification effectiveness, highlighting the need for robust segmentation strategies.

4.5. Discussion

The proposed segmentation-guided inpainting method demonstrated strong performance in recovering adversarially perturbed images. Key improvements were observed in preserving semantic integrity and classification accuracy. However, challenges remain in densely packed scenes with overlapping objects. Future work will explore hybrid approaches combining segmentation and object detection for further robustness.

5. Discussion and Conclusion

This work presents a segmentation-guided inpainting approach for adversarial purification, achieving significant improvements in precision and robustness compared to naive reconstruction methods. By leveraging segmentation masks to guide diffusion-based recovery, our method selectively addresses perturbed regions while preserving unaffected areas. Despite its strengths, several limitations remain:

Limitations: Our approach struggles with high noise levels, as the inpainting process becomes less reliable under strong perturbations. Not all inpainting pipelines are equally efficient, and balancing diffusion model capability with runtime remains a challenge. Additionally, FGSM-based attacks can unintentionally corrupt visual properties, complicating evaluation across different VLMs. Lastly, multiclass classification remains difficult to evaluate due to sparse target vectors, so a different surrogate for VLM effectiveness might be a point to research.

Future Directions: Promising directions include extending this approach to multimodal tasks, such as Visual Question Answering, and evaluating robustness against targeted attacks. Improving pipeline efficiency through alternative segmentation and inpainting techniques is essential for scalability. Finally, creating a diverse adversarial dataset will enable broader benchmarking and further enhance robustness.

In summary, our results demonstrate that segmentation-guided inpainting is a modestly effective solution for adversarial purification. Future work should explore its application to multimodal scenarios and complex adversarial tasks, as well as more complex inpainting pipelines, paving the way for more resilient vision-language systems.

Student Name	Contributed Aspects	Details
Rodrigo B. Loza	Segmentation and Inpainting	Implemented the segmentation pipeline and inpainting process, generated and analyzed results, drafted the report, prepared slides, and participated in discussions.
Yilun Zhou	Dataset and Detection, Documentation	Implemented the noised dataset, set up segmentation/image detection pipeline, drafted the report, prepared slides, and participated in discussions.
Jiachun Zhang	Documentation and Presentation	Drafted and prepared slides, participated in discussions.

Table 2. Contributions of Team Members

6. Work Division

The division of work among team members is summarized in Table 2. Each member contributed to distinct aspects of the project, ensuring its successful completion through collaborative efforts.

References

[1] Lars Ankile et al. Denoising diffusion probabilistic models as a defense against adversarial attacks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. Diffusion defense.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. CW attack.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. FGSM attack.

[4] Puneet Gupta et al. Ciidefence: Defeating adversarial attacks by fusing class-specific image inpainting and image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. Class-specific inpainting defense.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. PGD attack.

[6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. DeepFool attack.

[7] Weili Nie et al. Diffusion models for adversarial purification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. Diffusion purification.

[8] Haibo Zhang et al. Versatile defense against adversarial attacks on image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8280–8289, 2019. Versatile defense framework.