Pattern Recognition Letters 194 (2025) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Robust image classification with multi-modal large language models

Francesco Villani [a], Igor Maljkovic [a,1], Dario Lazzaro [a,b,1], Angelo Sotgiu [c], Antonio Emanuele Cinà [a,*], Fabio Roli [a]

[a] University of Genoa, Via Dodecaneso, 35, Genoa, 16146, Italy
[b] Sapienza University of Rome, Via Ariosto 25, Rome 00185, Italy
[c] University of Cagliari, Via Marengo 3, Cagliari 09100, Italy

ARTICLE INFO

MSC:
68T01
68T45
68T10
68T05

Keywords:
Adversarial machine learning
Robust classification
Multimodal large language model
Multimodal information
Adversarial examples
Machine learning security
TrustworthyAI

ABSTRACT

Deep Neural Networks are vulnerable to adversarial examples, i.e., carefully crafted input samples that can cause models to make incorrect predictions with high confidence. To mitigate these vulnerabilities, adversarial training and detection-based defenses have been proposed to strengthen models in advance. However, most of these approaches focus on a single data modality, overlooking the relationships between visual patterns and textual descriptions of the input. In this paper, we propose a novel defense, *Multi-Shield*, designed to combine and complement these defenses with multimodal information to further enhance their robustness. *Multi-Shield* leverages multimodal large language models to detect adversarial examples and abstain from uncertain classifications when there is no alignment between textual and visual representations of the input. Extensive evaluations on six distinct datasets, using robust and non-robust image classification models, demonstrate that *Multi-Shield* can be easily integrated to detect and reject adversarial examples, outperforming the original defenses.

1. Introduction

Deep Neural Networks (DNNs) are employed in numerous image recognition tasks [1], such as facial recognition, medical imaging, pedestrian segmentation for autonomous driving, and surveillance. For instance, in medical imaging, DNNs assist in detecting diseases like cancer from X-rays and MRIs [2]. In autonomous driving [3], DNNs help recognize pedestrians and road signs, improving safety and navigation. Their remarkable performance has, therefore, led to widespread integration into various applications, impacting user safety, daily routines, services, and product quality. However, these models are vulnerable to adversarial examples [4–6]—subtle, often imperceptible perturbations in the input data that can mislead the model into making incorrect predictions or classifications. Adversarial examples are typically generated through gradient-based evasion attacks [7,8], which exploit the model's sensitivity to minor input changes to induce erroneous predictions, often with high confidence. Furthermore, adversarial examples pose a significant threat to the reliability of DNNs, especially in user-facing and safety-critical applications. These attacks not only degrade model performance but also introduce serious concerns about compliance with emerging regulatory standards. For instance, frameworks like the AI Act [9] have imposed stricter demands on the reliability and accountability of AI systems. Therefore, as DNNs deployment expands and regulatory requirements tighten, developing robust defenses against adversarial examples is increasingly important for safe and responsible use.

Consequently, various defensive strategies have been proposed in recent years to address these challenges. Prominent among them are adversarial training and detectors, complementary approaches both representing proactive defenses that align with the *security-by-design* principle by incorporating security measures during model development [10]. Adversarial training [11,12] strengthens models by injecting knowledge of adversarial attacks during their training (e.g., including adversarial examples in the training data), improving their resilience to future attacks. Detector-based defenses, on the other hand, encompass techniques that detect and mitigate/reject attacks as they occur. Examples include anomaly detection systems that reject potentially dangerous inputs in real-time and abstain from predicting them [13–15]. However, each of these defense strategies has its own limitations. Adversarial training, while effective, can be complex to develop and resource-intensive [16], requiring continuous adaptation or retraining

* Corresponding author.
E-mail addresses: francesco.villani@edu.unige.it (F. Villani), igor.maljkovic@edu.unige.it (I. Maljkovic), dario.lazzaro@uniroma1.it (D. Lazzaro), angelo.sotgiu@unica.it (A. Sotgiu), antonio.cina@unige.it (A.E. Cinà), fabio.roli@unige.it (F. Roli).
[1] Authors contributed equally.

https://doi.org/10.1016/j.patrec.2025.04.022
Received 26 September 2024; Received in revised form 2 April 2025; Accepted 17 April 2025
Available online 3 May 2025

to address evolving attacks. Conversely, detectors rely on recognizing deviations from normal behavior, which may not be effective against new attacks [10].

To address these challenges, we propose *Multi-Shield*, a solution that integrates the strengths of both adversarial training and detection defenses while handling multimodal information. The idea is to combine adversarial training defenses with a rejection module that incorporates both visual and textual information during prediction. In particular, *Multi-Shield* uses an adversarially trained model and a multimodal model (i.e., CLIP [17]), which analyzes visual and textual patterns to verify their alignment. Specifically, the multimodal model uses visual images and textual descriptions related to the classes in the dataset to assess semantic agreement, employing zero-shot predictions. As illustrated in Fig. 1, when there is an agreement between the proactively trained image classifier and the multimodal model, the system provides a prediction. However, if there is a discrepancy, the system abstains from responding, as the lack of agreement suggests potential adversarial manipulation or uncertainty. Our experiments demonstrate the effectiveness of *Multi-Shield* on six benchmark datasets (i.e., CIFAR-10, ImageNet, Caltech-101, Food-101, Oxford-IIIT Pets, STL-10). We applied *Multi-Shield* to both non-robust models and robust models developed with adversarial training. The results reveal that *Multi-Shield* significantly increases robustness for proactively trained models, as it successfully rejects adversarial examples that would otherwise bypass them. Overall, these results highlight *Multi-Shield*'s ability to complement existing defenses and provide an additional layer of security for them.

In this work, we make the following key contributions:

- We propose *Multi-Shield*, a novel defense mechanism that leverages multimodal data, specifically visual and textual inputs, to identify and reject adversarial examples;
- We explore the interaction between adversarial training and a multimodal detector, demonstrating how *Multi-Shield* can be effectively integrated as a supplementary layer to enhance overall model robustness;
- We assess the robustness improvements provided by multimodal information against adversarial attacks by evaluating *Multi-Shield* on six datasets and thirteen models.

The rest of the paper is structured as follows: Section 2 provides an overview of adversarial defenses and the principles of multimodal large language models. In Section 3, we present the architecture and inner workings of *Multi-Shield*. Section 4 outlines our experimental setup and results, and Section 5 concludes with limitations and future research directions.

2. Background and related work

In this section, we present background information and review related work on adversarial attacks and defenses, as well as multimodal large language models.

2.1. Adversarial examples

Adversarial examples are carefully crafted inputs designed to mislead models into making incorrect predictions, with the purpose of revealing and exploiting vulnerabilities in machine learning models. Given an input sample $\boldsymbol{x} \in \mathcal{X} = [0,1]^d$ with its true label $y \in \mathcal{Y} = \{1,\dots,K\}$ (where K is the number of classes) and a trained classifier $f : \mathcal{X} \to \mathcal{Y}$, an adversarial example $\boldsymbol{x}'$ is a slightly perturbed version of $\boldsymbol{x}$, designed to cause misclassification. Formally, the adversarial example can be found by solving the following constrained optimization problem:

$$\min_{\boldsymbol{x}'} \; L(\boldsymbol{x}', y) \tag{1}$$

$$\text{s.t. } \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \varepsilon \tag{2}$$

$$\boldsymbol{x}' \in [0,1]^d, \tag{3}$$

where $L(\cdot)$ in Eq. (1) is a loss function that penalizes correct classification of $\boldsymbol{x}'$ by f. A commonly used loss function is the Difference of Logits [6], expressed as:

$$L(\boldsymbol{x}, y) = f_y(\boldsymbol{x}) - \max_{j \in \mathcal{Y} \setminus \{y\}} f_j(\boldsymbol{x}), \tag{4}$$

with $f_i(\boldsymbol{x})$ denoting the logit score of $\boldsymbol{x}$ associated with class i. Under this loss, the optimization program aims to find an adversarial example $\boldsymbol{x}'$ that decreases the logit corresponding to the true class label y while increasing the logits of competing incorrect classes. As a result, the model is driven to misclassify the input by predicting the incorrect class with the highest logit score. The constraint in Eq. (2) ensures that the adversarial example $\boldsymbol{x}'$ remains similar to the original input $\boldsymbol{x}$ by limiting the norm of their difference, with typical choices for $p \in \{0, 1, 2, \infty\}$ [6]. Lastly, the constraint in Eq. (3) ensures that $\boldsymbol{x}'$ remains within the valid input space $[0,1]^d$, a typical requirement for image recognition tasks. Given the above formulation, several optimization methods have been proposed to solve this problem, including PGD [18], FGSM [11], and many others [8]. Among them, AutoAttack [7] stands out as a state-of-the-art approach due to its ability to generate adversarial examples effectively without parameter tuning, making it well-suited for reliable and consistent benchmarking of model robustness.

2.2. Adversarial defenses

Prominent proactive defenses against adversarial examples in DNNs fall into two main categories: robust optimization techniques and adversarial detectors [10,15]. Among the most effective methods in the first category is adversarial training [11,12,19], which strengthens robustness by exposing the model to adversarially perturbed inputs during the training process and has consistently proven to be one of the most effective defense mechanisms to date [7,12]. Detectors, on the other hand, focus on mechanisms that enable the model to abstain from making predictions when it suspects inputs to be adversarial. The pioneering approach by [20] involves rejecting samples whose feature representations are significantly distant from class centroids. Similarly, [14] propose a distance-based rejection method using thresholds applied to outputs of a multi-class SVM to estimate data feature representation distribution. Later works extend the inspection of feature representations beyond the final layer of the network. [13] leverage late-stage ReLU activations combined with a quantized RBF-SVM to detect adversarial examples, though this method requires adversarial training, increasing computational cost. Alternatively, [21] present a KNN-based method that evaluates intermediate layer representations and rejects inputs with inconsistencies, though it incurs substantial overhead due to the need for distance computations with respect to training data at each layer. [15] propose a method for evaluating the consistency of feature representations across different layers, using a threshold for distance-based rejection.

However, all these strategies are restricted to single-modality data. They focus solely on the visual semantics of the input during training, overlooking the semantical alignment between predicted classes and visual patterns. Furthermore, the interaction between adversarial training and rejection remains unexplored, preventing the development of a more robust defense.

2.3. Multimodal vision-language models

Multimodal models represent a recent development in AI, enabling the simultaneous processing of multiple types of data [22]. These models are trained to process data from different modalities, identifying patterns aligning the information. In particular, vision-language map images and textual descriptions into a shared semantic space,

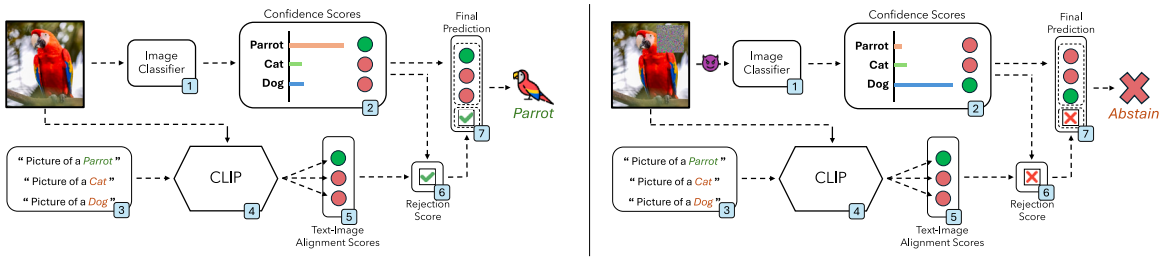

Fig. 1. Illustration of *Multi-Shield*'s operation. An image classifier (1) first processes the input image to generate an initial prediction (2). Simultaneously, the CLIP model (4) performs zero-shot classification using text prompts (3). A rejection score (6) is computed by comparing the agreement between the image classifier's prediction (2) and the CLIP model's output (5). Lastly, *Multi-Shield* either returns the final prediction or abstains (7), according to the uncertainty in classification. The left part illustrates a clean prediction, while the right part shows *Multi-Shield* abstaining for an adversarial example.

facilitating tasks that require a joint understanding of both modalities [23]. Applications of these models include image captioning [24], visual question answering [25], and text-to-image retrieval [26], where the combination of visual and textual information allows models to better capture contextual meaning and improve performance. In this work, we focus on vision-language models applied to image recognition tasks. Several prominent vision-text multimodal models have emerged, including VisualBERT [27], VilBERT [28], and VisionLLaMA [29], each designed to capture complex relationships between visual and textual inputs. These models typically employ transformer-based architectures [30] to encode visual and textual data into a unified representation. Another notable example of vision-text multimodal is CLIP (Contrastive Language-Image Pretraining) [17], which aligns visual and textual embeddings using large-scale datasets of paired images and captions. CLIP works by training on a contrastive learning objective, ensuring that semantically similar images and text are positioned closely within a shared embedding space, leading to strong performance in tasks such as zero-shot image classification and text-to-image retrieval [17].

3. *Multi-shield*'s rejection mechanism

We now present the *Multi-Shield* structure, detailing its operations, problem formulation, and the design of an adaptive attack to evaluate its robustness under worst-case scenarios.

3.1. Multishield's workflow

The *Multi-Shield* framework operates in three distinct phases, depicted in Fig. 1, each contributing to a robust image classification process that leverages both unimodal and multimodal analysis. We name these phases with: *unimodal prediction*, *multimodal alignment*, and *Multi-Shield decision*.

Unimodal Prediction. The *Multi-Shield* workflow begins with the image classifier processing the input (Step 1) and generating confidence scores for each class in the dataset (Step 2). Based on these scores, an initial prediction is made, relying solely on the image classifier's output. Notably, the image classifier at this stage can be either a standard (non-robust) model or a robust model trained to resist adversarial attacks. In either case, *Multi-Shield* operates on top of the classifier, augmenting its defense with an additional verification layer.

Multimodal Alignment. In this second phase, *Multi-Shield* uses the CLIP model as a zero-shot vision-language classifier [17]. It compares the visual representation of the input image with class description prompts to determine alignment. For each class, we automatically create natural language prompts in the format `'Picture of a [object]'` (Step 3). For example, for classes such as 'parrot' and 'dog', the prompts would be `'Picture of a parrot'` and `'Picture of a dog'`, respectively. The text prompts serve as a simplified but effective method to represent each class conceptually; however, with some human effort, they could be further enriched by including additional descriptive characteristics (e.g., mentioning that parrots have feathers and wings). During prediction, these prompts, along with the input image, are processed through CLIP's dual encoders: one for images and one for text (Step 4). This process generates visual and textual embeddings that capture the features of both modalities, with CLIP aiming to maximize the alignment between the image and the textual description that best matches it. Finally, *Multi-Shield* performs zero-shot classification with CLIP by computing alignment scores using cosine similarity and selecting the class with the highest score as the final prediction [17] (Step 5).

Multi-Shield Decision. In the final phase, *Multi-Shield* compares the predictions from the unimodal image classifier and the multimodal CLIP model (Step 6), each capturing distinct patterns in the data. If both models agree, *Multi-Shield* outputs their shared prediction; if not, it abstains (Step 7). As shown on the left side of Fig. 1, for a clean image, both models confidently identify the class 'parrot', leading *Multi-Shield* to output the agreed-upon prediction. Conversely, on the right side of Fig. 1, adversarial noise causes the unimodal classifier to misclassify the image, while the CLIP model still correctly aligns the image with the 'parrot' class. This disagreement drives *Multi-Shield* to abstain, acknowledging that the adversarial input may have misled one model. Unlike typical robust models, which do not employ a rejection mechanism, *Multi-Shield* introduces this feature to abstain when uncertain (e.g., input is suspected to be adversarial) and ensures predictions are made only when both models agree.

3.2. Problem formulation

Multi-Shield is defined as a classification function $\hat{f} : \mathcal{X} \to \hat{\mathcal{Y}}$, where $\hat{\mathcal{Y}} = \mathcal{Y} \cup \{K+1\}$ denotes the set of class labels in the dataset $\mathcal{Y}$ extended with an additional *rejection class*. At inference time, for each input sample $\boldsymbol{x} \in \mathcal{X}$, *Multi-Shield* computes a rejection score $R(\boldsymbol{x})$, which determines whether to make a prediction or abstain. The final decision of $\hat{f}$ is then given by:

$$\hat{f}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & \text{if } R(\boldsymbol{x}) \leq 0 \\ k+1, & \text{if } R(\boldsymbol{x}) > 0. \end{cases} \tag{5}$$

Here, $R(\boldsymbol{x})$ represents the level of agreement between the unimodal image classifier and the zero-shot multimodal prediction from CLIP, serving as a score to gauge the model's uncertainty about the input. When $R(\boldsymbol{x})$ is non-positive, it indicates that the image classifier and CLIP are in agreement, suggesting high confidence in the prediction. Conversely, if $R(\boldsymbol{x})$ is positive, the model abstains from making a prediction due to uncertainty. The rejection score $R(\boldsymbol{x})$ of *Multi-Shield* is computed as:

$$R(\boldsymbol{x}) = \left| \max_{i \in \{1,\ldots,K\}} h(\boldsymbol{x}, P_i) - h(\boldsymbol{x}, P_j) \right|, \tag{6}$$

where $j = f(\boldsymbol{x})$ is the predicted class from the image classifier and P_i represents the textual prompt for class i. The function h measures

the text-image alignment scores using CLIP's zero-shot capabilities, computing the cosine similarity between the embeddings of the image and the textual prompts in a shared semantic space. Formally:

$$h(\boldsymbol{x}, P_i) = \frac{h_{\text{img}}(\boldsymbol{x}) \cdot h_{\text{txt}}(P_i)}{|h_{\text{img}}(\boldsymbol{x})||h_{\text{txt}}(P_i)|} \tag{7}$$

where $h_{\text{img}}(\boldsymbol{x}) \in \mathbb{R}^d$ is the visual embedding of the input image and $h_{\text{txt}}(P_i) \in \mathbb{R}^d$ is the textual embedding for the class prompt i, both generated using the encoders of the multimodal model. *Multi-Shield* evaluates the alignment score for each class prompt and picks the highest one to perform zero-shot classification using the CLIP model. Finally, when *Multi-Shield* abstains, the positive score in $R(\boldsymbol{x})$ indicates the level of disagreement between the image classifier and CLIP's classification, supporting decisions on abstention reliability and subsequent input handling.

3.3. Attacking multi-shield

To evaluate the worst-case robustness of *Multi-Shield*, we assess its performance against an adaptive attacker [31], who is fully aware of the defense and actively seeks to circumvent it. Adaptive attacks indeed provide a more rigorous and stronger test of defenses [15,21,32]. The attacker, having access to the image recognition model and the rejection mechanism in *Multi-Shield*, aims to simultaneously deceive the unimodal image classifier and manipulate the zero-shot multimodal classifier, causing both to agree on an incorrect label. The adaptive attack can thus be formalized by revisiting the optimization program in Eqs. (1)–(3) as follows:

$$\min_{\boldsymbol{x}'} L(\boldsymbol{x}', y) \tag{8}$$

$$\text{s.t. } \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \varepsilon \tag{9}$$

$$\boldsymbol{x}' \in [0,1]^d \tag{10}$$

$$R(\boldsymbol{x}') \leq 0 \tag{11}$$

The additional constraint in Eq. (11) ensures that the adversarial input does not trigger the rejection mechanism of *Multi-Shield*.

4. Experiments

In this section, we evaluate the effectiveness of *Multi-Shield*'s rejection mechanism in defending against adversarial attacks. Our goal is to assess *Multi-Shield*'s performance under both traditional (non-adaptive) and adaptive attack scenarios.

4.1. Experimental setup

Datasets. We consider six datasets: CIFAR-10 [33], ImageNet [34], Caltech-101 [35], Food-101 [36], Oxford-IIIT Pets [37], STL-10 [38]. For CIFAR-10, we use all 10,000 test images. For ImageNet, we randomly sample 5000 validation images. We evaluate the full validation sets for Caltech-101, Oxford-IIIT Pets, and STL-10, and sample 10000 images for Food-101.

Classifiers. For the image classifier, we source baseline and top state-of-the-art ℓ_∞ robust models from the RobustBench library [39]. For CIFAR-10, we evaluate six models denoted with `C1`-`C6`. `C1` [39] is a non-robust model. `C2` [40], `C4` [41] and `C5` [42] employ adversarial training combined with data augmentation. `C3` [43] utilizes a robust ensemble of models, while `C6` [44] reformulates adversarial training to maximize the margin for more vulnerable samples, enhancing overall robustness. For ImageNet, we select four robust models, denoted with `I1`-`I4`, incorporating adversarial training. `I1` [45] and `I4` [46] utilize a ViT architecture, `I2` [47] employs a Swin Transformer, and `I3` [46] uses the ConvNeXt architecture. For experiments on Caltech-101, Oxford-IIIT Pets, STL-10, and Food-101, we fine-tune three ImageNet-pretrained models—ResNet-50 (`M1`), WideResNet50-2 (`M2`), and ViT-B/16 (`M3`). For each model, we replace the final classification layer and fine-tune the entire model for each dataset over 50 epochs using stochastic gradient descent with a learning rate of 0.001 and momentum of 0.9.

Multishield Construction. The *Multi-Shield* module combines a CLIP model with a standard image classifier. For CIFAR-10, we use a ViT-B visual encoder fine-tuned for CIFAR-10 classification [48], and for ImageNet, a pre-trained ViT-L [49], both paired with their respective text encoders from Hugging Face [50].

Attack. We conduct the experiments using AutoAttack [7], a state of the art algorithm. We use both a naive (non-adaptive) version of AutoAttack and an adaptive variant, which takes into account *Multi-Shield*'s rejection mechanism. Lastly, we set AutoAttack's perturbation size to $\varepsilon = 8/255$, as typical reference for robustness evaluation [39].

Evaluation Metrics. In our experiments, we assess *Multi-Shield*'s performance using several key metrics. Clean Accuracy measures the accuracy on unperturbed inputs, providing a baseline for the model's performance on pristine data. In this case, rejected samples are considered errors. The difference between the Clean Accuracy without any defense in place and *Multi-Shield*'s Clean Accuracy corresponds to the fraction of wrongly rejected samples (i.e., false positives). Robust Accuracy evaluates how well *Multi-Shield* performs on adversarial inputs by accounting for correct predictions and cases where it abstains from making a prediction. In fact, the desired model behavior in the presence of adversarial examples is to avoid misclassification. The Rejection Ratio represents the percentage of inputs for which *Multi-Shield* opts to abstain, reflecting the effectiveness of the rejection mechanism in identifying adversarial examples. The difference between the Robust Accuracy and Rejection Ratio corresponds to the fraction of correctly classified test inputs. Lastly, all experiments have been run on a NVIDIA A100 Tensor Core GPU with 40 GB of VRAM.

4.2. Experimental results

We present in Table 1 the results on the effectiveness of *Multi-Shield* in detecting adversarial examples across three distinct attack scenarios: (1) the baseline image classifiers is the target of attack without *Multi-Shield* detection (Baseline); (2) *Multi-Shield* is integrated during the final prediction while the image classifier is under attack (Multi-Shield); and (3) an adaptive attacker targets both the image classifier and *Multi-Shield*, providing a worst-case evaluation of the defense mechanism (Multi-Shield - Adaptive Attack). We then integrate our analysis with Fig. 2 showing the performance of *Multi-Shield* against adversaries of varying strengths ε. Lastly, we compare *Multi-Shield* with two state-of-the-art rejection defenses (i.e., [14,15]) in Table 2.

***Multi-Shield* Effectiveness.** We analyze the robustness of pre-trained models without *Multi-Shield* (Baseline) and assess the impact of integrating *Multi-Shield* to enable rejection of anomalous samples (Multi-Shield). The results in Table 1 confirm that *Multi-Shield* serves as an effective additional security layer significantly enhancing robust accuracy across all models without requiring their retraining. For example, *Multi-Shield* raises the robust accuracy of non-robust models (i.e., `C1`, `M1`, `M2`, `M3`) from near 0% (Baseline) to over 90% across multiple datasets, effectively rejecting most adversarial examples. The robustness boost with *Multi-Shield* is notable even on robust models, with an average improvement of 32% on CIFAR-10 and 65% on ImageNet (Multi-Shield). The only trade-off is a slight reduction in clean accuracy, averaging approximately 1.8% on CIFAR-10, 6% on ImageNet and Caltech-101, 3% on Food-101 and Oxford-IIIT Pets, and under 1% on STL-10.

Impact of Attack Strength on *Multi-Shield*. We evaluate *Multi-Shield* against attackers of varying strengths (i.e., at different values of ε), demonstrating how robust accuracy and rejection ratio evolve as attack

Table 1
Experimental results on baseline and adversarially pre-trained models (6 on CIFAR-10, 4 on ImageNet, and 3 on Caltech-101, Food-101, Oxford-IIIT Pets, and STL-10) evaluated using clean accuracy (%), robust accuracy (%), and rejection ratio (%). The defense is assessed using AutoAttack with perturbation size $\epsilon = 8/255$ across three defense scenarios: without *Multi-Shield* (Baseline), with *Multi-Shield* (Multi-Shield), and adaptively attacking *Multi-Shield* (Multi-Shield - Adaptive Attack). Note that performing the attack on either of the underlying clip models yields a robust accuracy of 0%.

Models	Baseline		Multi-Shield			Multi-Shield - Adaptive Attack	
	Clean Acc.	Robust Acc.	Clean Acc.	Robust Acc.	Rejection Ratio	Robust Acc.	Rejection Ratio
CIFAR-10							
C1	94.8	0.0	92.7	91.8	91.8	20.4	0.4
C2	89.7	59.9	87.8	92.0	32.5	63.6	0.7
C3	86.1	52.0	84.5	91.9	40.3	55.0	0.4
C4	88.7	66.8	87.0	93.3	27.2	69.6	0.8
C5	85.7	52.9	84.2	92.2	39.5	58.4	0.5
C6	93.7	65.6	91.6	88.4	23.4	66.9	0.6
ImageNet							
I1	70.9	10.3	63.8	88.7	78.7	17.4	4.4
I2	75.1	27.0	67.1	89.4	63.5	32.5	2.8
I3	74.9	26.2	67.3	87.8	62.6	30.7	3.0
I4	74.7	22.7	66.7	87.9	66.0	28.4	2.7
Caltech-101							
M1	97.2	0.3	90.5	97.1	96.8	71.3	4.2
M2	97.4	0.2	90.9	97.7	97.6	67.3	3.3
M3	96.5	0.0	90.1	94.8	94.8	39.2	4.6
Food-101							
M1	80.3	0.0	77.3	97.7	97.7	74.4	26.2
M2	82.6	0.0	79.4	96.9	96.9	65.4	21.1
M3	86.2	0.0	83.4	98.6	98.6	27.9	12.4
Oxford-IIIT Pets							
M1	93.1	1.1	89.6	94.8	93.8	28.4	6.8
M2	93.3	1.1	89.8	94.5	93.4	22.2	6.1
M3	91.3	0.5	88.2	97.8	97.2	18.4	6.5
STL-10							
M1	97.9	0.3	97.2	98.8	98.5	61.2	0.5
M2	98.2	0.4	97.6	98.4	97.9	51.5	0.3
M3	97.8	0.0	97.0	99.2	99.2	17.2	0.1

Table 2
Comparison of *Multi-Shield*, NR [14], and DNR [15] over five runs on 1,000-sample subsets from CIFAR-10. The rejection thresholds for NR and DNR are set to roughly match *Multi-Shield*'s accuracy. *Robust Acc. (A)* refers to accuracy under an adaptive attack.

Defense	Model	Clean acc.	Robust acc.	Robust acc. (A)	Model	Clean acc.	Robust acc.	Robust acc. (A)
NR		91.7 ± 0.002	3.2 ± 0.001	0.00 ± 0.00		88.4 ± 0.01	75.6 ± 0.03	59.2 ± 0.04
DNR	C1	91.5 ± 0.001	**96.1 ± 0.004**	0.00 ± 0.00	C2	88.2 ± 0.006	76.3 ± 0.02	60.6 ± 0.05
Ours		92.7 ± 0.002	91.9 ± 0.006	**20.2 ± 0.003**		88.3 ± 0.007	**91.7 ± 0.009**	**64.2 ± 0.05**
NR		84.3 ± 0.02	72.1 ± 0.02	49.3 ± 0.05		90.7 ± 0.007	75.9 ± 0.03	66.3 ± 0.04
DNR	C5	84.2 ± 0.009	73.1 ± 0.04	50.8 ± 0.04	C6	90.8 ± 0.009	76.0 ± 0.03	66.6 ± 0.03
Ours		83.8 ± 0.004	**92.5 ± 0.008**	**58.9 ± 0.05**		91.3 ± 0.002	**88.8 ± 0.009**	**66.9 ± 0.07**
NR	C3	82.6 ± 0.003	66.1 ± 0.01	43.8 ± 0.03	C4	82.3 ± 0.006	78.7 ± 0.03	62.1 ± 0.05
Ours		84.5 ± 0.003	**91.9 ± 0.02**	**55.0 ± 0.03**		87.0 ± 0.002	**93.3 ± 0.008**	**69.6 ± 0.06**

power increases. These trends are illustrated in Fig. 2 across four robust models (C2,C6,I2,I4) on CIFAR-10 (top two plots) and ImageNet (bottom two plots). Firstly, in all three plots, the gap between the robust accuracy of *Multi-Shield* (solid blue line) and the baseline (dashed gray line) widens as ϵ increases, highlighting *Multi-Shield*'s superior robustness in response to increasing attack strength. Secondly, the rejection ratio (red line) steadily increases in all three models, showcasing *Multi-Shield*'s ability to detect and reject stronger adversarial examples that mislead the baseline. These results become more evident in the ImageNet models (Figs. 2(c)–2(d)), where robust accuracy remains relatively more resilient while rejections rise significantly.

Adaptive Attacker. We conclude our analysis by evaluating *Multi-Shield* under the worst-case scenario of an adaptive attacker who knows everything about the target model and exploits knowledge of the defense mechanism to bypass it [32]. Specifically, we give the strong assumption of an adaptive attacker with complete knowledge of the rejection mechanism, including *Multi-Shield*'s image classifier, rejection mechanism, and inner workings (i.e., parameters and class prompts) to have a complete system evaluation. Firstly, looking at the rightmost scenario in Table 1 (Multi-Shield - Adaptive Attack), *Multi-Shield* provides substantial improvements in robust accuracy across all non-robust models (C1, M1, M2, M3). While the gains are already significant for CIFAR-10 (20.4%) and Oxford-IIIT Pets (22.1%), they become even larger for Caltech-101 (59.1%), Food-101 (55.9%), and STL-10 (43%). For robust models, *Multi-Shield* enhances robustness by an average of 3.3% on CIFAR-10 and 5% on ImageNet. This indicates that attacking *Multi-Shield*, even under worst-case evaluation, requires significantly more effort and introduces further difficulties for the attacker. The constraint in Eq. (11) forces the attack to simultaneously deceive both the unimodal image classifier and the multimodal CLIP classifier, complicating the optimization process. We conclude that *Multi-Shield* serves as an additional layer of defense to enhances the robustness of all considered models, both under non-adaptive and adaptive (worst-case) attack settings, across all datasets.

Comparison with other Rejection Defenses. We compare our method with state-of-the-art rejection defenses, including Neural Rejection

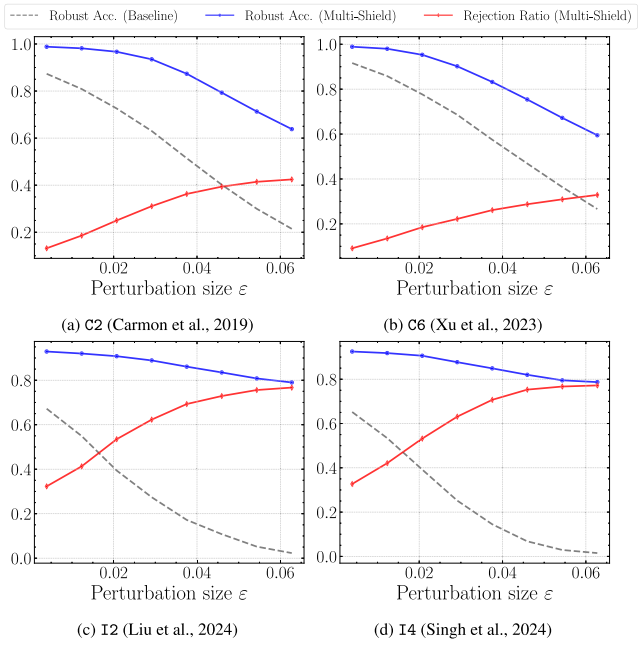

Fig. 2. Curves showing the baseline robust accuracy alongside *Multi-Shield*'s robust accuracy and rejection ratio under attacks with varying perturbation sizes (ε) across different robust models on CIFAR-10 (top two plots) and ImageNet (bottom two plots).

(NR) [14] and Deep Neural Rejection [15], following the experimental setups outlined in their original works. Experiments are conducted on five distinct subsets of 1000 CIFAR-10 samples, reporting the average and standard deviation of results. The results, reported in Table 2, showcase that *Multi-Shield* consistently outperforms NR and DNR in robustness across various models. Notably, while DNR achieves high robustness under C1 (e.g., 96.1%), it fails entirely under adaptive attacks, dropping to 0.00%. In contrast, *Multi-Shield* maintains significantly higher robustness against adaptive attackers (e.g., 20.2% vs. 0.00% for C1). Similar improvements are observed across other robust models, and the low standard deviation across multiple runs confirms the consistency of our results, reinforcing the validity of our conclusions. On C5, *Multi-Shield* achieves 58.9% robust accuracy under adaptive attacks, outperforming NR and DNR by approximately 8 and 9 percentage points, respectively. Likewise, on C4, *Multi-Shield* reaches 69.6% robust accuracy, surpassing NR (62.1%) by 7.5%. Beyond robustness, *Multi-Shield* is also computationally less demanding. Both NR and DNR require training one or multiple one-vs-all SVMs for each class, leading to substantial memory and computational overhead. DNR, in particular, necessitates extreme batch size reductions (e.g., batch size = 1) to fit within 40 GB of VRAM when applied to CIFAR-10, making it impractical for larger datasets (e.g., ImageNet) or more complex models (e.g., C3, C4). We thus conclude that these comparisons further emphasize the benefits of *Multi-Shield* as an additional security layer for robust classification systems, enhancing resilience against adversarial threats while maintaining high accuracy and computational efficiency.

5. Conclusions, limitations, and future works

In this work, we introduce *Multi-Shield*, a novel defense that implements a rejection mechanism based on two core principles: (i) integration of multimodal information and (ii) interaction between adversarial training and a multimodal adversarial detector. The first principle is realized by analyzing visual features and their semantic alignment with the textual prompts for the classes. In this way, *Multi-Shield* enables the identification and rejection of adversarial examples when classification presents uncertainties or semantic inconsistencies. It leverages agreement in predictions between an image classifier and a multimodal model to decide whether to abstain from making a prediction. Secondly, *Multi-Shield* complements existing defenses that rely on adversarial training, providing an additional security layer against adversarial attacks. Our experiments show that *Multi-Shield* consistently detects adversarial examples that mislead image classifiers and remains effective even under worst-case adaptive attacks. Despite its strengths, *Multi-Shield* may struggle to detect adversarial inconsistencies in datasets with abstract or meaningless labels, such as "Class 1" or "Class 2", as these provide no meaningful semantic cues for the multimodal model to align textual and visual data. For future work, we aim to explore alternative methods for crafting more descriptive class label prompts, investigate their impact on robustness, and extend adversarial training to the multimodal model to further strengthen the defense.

CRediT authorship contribution statement

Francesco Villani: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation. **Igor Maljkovic:** Visualization, Validation, Software. **Dario Lazzaro:** Writing – review & editing, Software, Methodology. **Angelo Sotgiu:** Writing – review & editing, Methodology. **Antonio Emanuele Cinà:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Fabio Roli:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This work has been partially supported by the EU-NGEU National Sustainable Mobility Center (CN00000023), Italian Ministry of University and Research Decree n. 1033-17/06/2022 (Spoke 10), and project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union-NextGenerationEU. Lastly, this work was carried out while Dario Lazzaro was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome in collaboration with the University of Genoa.

Acknowledgments

This work has been partially supported by the EU—NGEU National Sustainable Mobility Center (CN00000023), Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10), and project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union—NextGenerationEU. Lastly, this work was carried out while Dario Lazzaro was enrolled in the Italian National Doctorate on Artificial Intelligence run by the Sapienza University of Rome in collaboration with the University of Genoa.

Data availability

No data was used for the research described in the article.

References

[1] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. 29 (9) (2017) 2352.
[2] P.K. Mall, P.K. Singh, S. Srivastav, V. Narayan, M. Paprzycki, T. Jaworska, M. Ganzha, A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities, Heal. Anal. (2023) 100216.
[3] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE Access (2020) 58443–58469.

[4] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: ECML PKDD, vol. 8190, 2013, pp. 387–402.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I.J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations, 2014.

[6] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy, 2017, pp. 39–57.

[7] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, in: International Conference on Machine Learning, 2020, pp. 2206–2216.

[8] A.E. Cinà, J. Rony, M. Pintor, L. Demetrio, A. Demontis, B. Biggio, I.B. Ayed, F. Roli, AttackBench: Evaluating gradient-based attacks for adversarial examples, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, http://dx.doi.org/10.1609/aaai.v39i3.32263.

[9] T. Madiega, Artificial intelligence act, Eur. Parliam.: Eur. Parliam. Res. Serv. (2021).

[10] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, in: ACM SIGSAC Conf. on Computer and Communications Security, 2018, pp. 2154–2156.

[11] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.

[12] A. Madry, Towards deep learning models resistant to adversarial attacks, 2017, arXiv:1706.06083.

[13] J. Lu, T. Issaranon, D. Forsyth, Safetynet: Detecting and rejecting adversarial examples robustly, in: IEEE International Conference on Computer Vision, 2017, pp. 446–454.

[14] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, F. Roli, Is deep learning safe for robot vision? adversarial examples against the icub humanoid, in: IEEE International Conference on Computer Vision Workshops, 2017.

[15] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, F. Roli, Deep neural rejection against adversarial examples, EURASIP J. Inf. Secur. 2020 (2020) 1–10.

[16] A. Shafahi, M. Najibi, M.A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L.S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, Adv. Neural Inf. Process. Syst. 32 (2019).

[17] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, pp. 8748–8763.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

[19] Z. Li, M. Wu, C. Jin, D. Yu, H. Yu, Adversarial self-training for robustness and generalization, Pattern Recognit. Lett. 185 (2024) 117–123.

[20] A. Bendale, T.E. Boult, Towards open set deep networks, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 1563–1572.

[21] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, 2018, arXiv:1803.04765.

[22] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, 2023, arXiv:2306.13549.

[23] Y. Du, Z. Liu, J. Li, W.X. Zhao, A survey of vision-language pre-trained models, 2022, arXiv:2202.10936.

[24] M.Z. Hossain, F. Sohel, M.F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Comput. Surv. 51 (6) (2019) 1–36.

[25] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. Van Den Hengel, Visual question answering: A survey of methods and datasets, Comput. Vis. Image Underst. 163 (2017) 21–40.

[26] M. Cao, S. Li, J. Li, L. Nie, M. Zhang, Image-text retrieval: A survey on recent research and development, 2022, arXiv:2203.14713.

[27] L.H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv:1908.03557.

[28] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Adv. Neural Inf. Process. Syst. 32 (2019).

[29] X. Chu, J. Su, B. Zhang, C. Shen, VisionLLaMA: A unified LLaMA interface for vision tasks, 2024, arXiv:2403.00522.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[31] N. Carlini, D. Wagner, Adversarial examples are not easily detected: Bypassing ten detection methods, in: ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3–14.

[32] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On evaluating adversarial robustness, 2019, arXiv preprint arXiv:1902.06705.

[33] A. Krizhevsky, Learning multiple layers of features from tiny images, 2009.

[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).

[35] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, in: Conference on Computer Vision and Pattern Recognition Workshop, 2004, pp. 178–178.

[36] L. Bossard, M. Guillaumin, L. Van Gool, Food-101–mining discriminative components with random forests, in: European Conference on Computer Vision, 2014, pp. 446–461.

[37] O.M. Parkhi, A. Vedaldi, A. Zisserman, C. Jawahar, Cats and dogs, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[38] A. Coates, A. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: International Conference on Artificial Intelligence and Statistics, 2011, pp. 215–223.

[39] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, RobustBench: a standardized adversarial robustness benchmark, in: Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.

[40] Y. Carmon, A. Raghunathan, L. Schmidt, J.C. Duchi, P.S. Liang, Unlabeled data improves adversarial robustness, in: Conf. on Neural Information Processing Systems, 2019.

[41] S. Gowal, S. Rebuffi, O. Wiles, F. Stimberg, D.A. Calian, T.A. Mann, Improving robustness using generated data, in: Advances in Neural Information Processing Systems, 2021, pp. 4218–4233.

[42] S. Addepalli, S. Jain, R. Venkatesh Babu, Efficient and effective augmentation strategy for adversarial training, in: Advances in Neural Information Processing Systems, 2022.

[43] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, Z. Wang, Adversarial robustness: From self-supervised pre-training to fine-tuning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 696–705.

[44] Y. Xu, Y. Sun, M. Goldblum, T. Goldstein, F. Huang, Exploring and exploiting decision boundary dynamics for adversarial robustness, in: International Conference on Learning Representations, 2023.

[45] E. Debenedetti, V. Sehwag, P. Mittal, A light recipe to train robust vision transformers, in: IEEE Conference on Secure and Trustworthy Machine Learning, 2023.

[46] N.D. Singh, F. Croce, M. Hein, Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models, Adv. Neural Inf. Process. Syst. 36 (2024).

[47] C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, S. Zheng, A comprehensive study on robustness of image classification models: Benchmarking and rethinking, Int. J. Comput. Vis. (2024) 1–23.

[48] A. Tang, L. Shen, Y. Luo, H. Hu, B. Do, D. Tao, FusionBench: A comprehensive benchmark of deep model fusion, 2024, arXiv preprint arXiv:2406.03280.

[49] X. Li, Z. Wang, C. Xie, CLIPA-v2: Scaling CLIP training with 81.1% zero-shot ImageNet accuracy within a $10, 000 budget; an extra $4, 000 unlocks 81.8% accuracy, 2023, arXiv, abs/2306.15658.

[50] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, HuggingFace's transformers: State-of-the-art natural language processing, 2019, arXiv, abs/1910.03771.