Towards Robust Vision Transformer via Masked Adaptive Ensemble

Fudong Lin
University of Delaware
Newark, DE, USA

Jiadong Lou
University of Delaware
Newark, DE, USA

Xu Yuan*
University of Delaware
Newark, DE, USA

Nian-Feng Tzeng
University of Louisiana at Lafeyette
Lafeyette, LA, USA

ABSTRACT

Adversarial training (AT) can help improve the robustness of Vision Transformers (ViT) against adversarial attacks by intentionally injecting adversarial examples into the training data. However, this way of adversarial injection inevitably incurs standard accuracy degradation to some extent, thereby calling for a trade-off between standard accuracy and adversarial robustness. Besides, the prominent AT solutions are still vulnerable to adaptive attacks. To tackle such shortcomings, this paper proposes a novel ViT architecture, including a detector and a classifier bridged by our newly developed adaptive ensemble. Specifically, we empirically discover that detecting adversarial examples can benefit from the Guided Backpropagation technique. Driven by this discovery, a novel Multi-head Self-Attention (MSA) mechanism is introduced for enhancing our detector to sniff adversarial examples. Then, a classifier with two encoders is employed for extracting visual representations respectively from clean images and adversarial examples, with our adaptive ensemble to adaptively adjust the proportion of visual representations from the two encoders for accurate classification. This design enables our ViT architecture to achieve a better trade-off between standard accuracy and adversarial robustness. Besides, the adaptive ensemble technique allows us to mask off a random subset of image patches within input data, boosting our ViT's robustness against adaptive attacks, while maintaining high standard accuracy. Experimental results exhibit that our ViT architecture, on CIFAR-10, achieves the best standard accuracy and adversarial robustness of 90.3% and 49.8%, respectively.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

KEYWORDS

Adversarial Training; Vision Transformer; AI Security

*Corresponding author: Dr. Xu Yuan (xyuan@udel.edu)

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679750

ACM Reference Format:
Fudong Lin, Jiadong Lou, Xu Yuan, and Nian-Feng Tzeng. 2024. Towards Robust Vision Transformer via Masked Adaptive Ensemble. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3627673.3679750

1 INTRODUCTION

The Vision Transformers (ViT) architecture has demonstrated impressive capabilities in a wide range of vision tasks, including image and video classification [2, 12, 73], dense prediction tasks [46, 80, 81], self-supervised learning [3, 20, 72], among others [6, 13, 24, 27, 36, 38, 39, 41, 49, 50, 62, 62, 89]. However, similar to Convolutional Neural Networks (CNNs) [21, 25, 28, 32, 40, 61, 63, 68–70, 92], the ViT architecture is vulnerable to adversarial attacks [5, 11, 18, 54–56, 91] achieved by maliciously altering clean images within a small distance, leading to incorrect predictions with high confidence. This vulnerability hinders the adoption of ViT in critical domains such as healthcare, finances, *etc.*

So far, adversarial training (AT) methods [1, 29, 30, 35, 52, 63, 79, 82, 90] are widely accepted as the most effective mechanisms for improving ViT's robustness against adversarial attacks, by intentionally injecting adversarial examples into the training data. Unfortunately, existing AT solutions struggle with two limitations. First, they suffer from a trade-off between standard accuracy (*i.e.*, the accuracy on clean images) and adversarial robustness (*i.e.*, the accuracy on adversarial examples), with improved robustness while yielding non-negligible standard accuracy degradation. Second, these solutions are not effective against adaptive attacks [9, 45, 75, 86], *i.e.*, a category of adversarial attacks capable of exploiting the weak points of defense methods to adaptively adjust their attack strategies. Hence, it calls for the exploration of enhancing ViT's robustness against adaptive attacks.

One potential direction to tackle the trade-off between standard accuracy and adversarial robustness is the *detection/rejection* mechanism. This involves training an additional detector to identify and reject malicious input data, with several solutions proposed in the literature [51, 58, 60, 74, 88]. However, these detection techniques have limited effectiveness against adaptive attacks and cannot be applied to scenarios involving natural adversarial examples, as reported in a prior study [26]. Hence, it is crucial to develop novel solutions that can address limitations associated with the aforementioned direction and are suitable for a wide range of scenarios.

In this work, we aim to boost the robustness of ViT against adaptive attacks in a more general and challenging scenario where

malicious inputs cannot be rejected. Such a scenario is common to several critical application domains, such as autonomous driving, where the system must correctly recognize a road sign even if it has been maliciously crafted. To this end, we propose a novel ViT architecture consisting of a detector and a classifier, connected by a newly developed *adaptive ensemble*. After adversarially trained by One-step Least-Likely Adversarial Training, our proposed ViT architecture can withstand adaptive attacks while incurring only a negligible standard accuracy degradation.

In essence, our detector incorporates two innovative designs to make adversarial examples more noticeable. First, based on our empirical observations, we introduce a novel Multi-head Self-Attention (MSA) mechanism [78] to expose adversarial perturbation by Guided Backpropagation [71]. Second, the Soft-Nearest Neighbors Loss (SNN Loss) [14, 64] is tailored to push adversarial examples away from their corresponding clean images. Our detector thus can effectively sniff adaptive attack-generated adversarial examples. On the other hand, our classifier's adversarial training involves two stages: pre-training and fine-tuning. During the pre-training stage, our classifier utilizes one clean encoder, one adversarial encoder, and one decoder to jointly learn high-quality visual representations and encourage pairwise similarity between a clean image and its adversarial example. Here, we extend Masked Autoencoders (MAE) [20] to facilitate adversarial training through a new design. Specifically, we reconstruct images from one pair of a masked clean image and its masked adversarial example, for representation learning, with a contrastive loss on a pair of visual representations to encourage similarity. In the fine-tuning stage, we discard the decoder and freeze the weights in the well-trained detector and two encoders, with a newly developed *adaptive ensemble* to bridge the detector and the two encoders, for fine-tuning an MLP (Multi-layer Perceptron) for accurate classification. Our adaptive ensemble also masks off a random subset of image patches within the input, enabling our approach to mitigate adversarial effects when encountering malicious inputs. Extensive experimental results on three popular benchmarks demonstrate that our approach outperforms state-of-the-art adversarial training techniques in terms of both standard accuracy and adversarial robustness.

2 RELATED WORK

Detection Mechanisms. Detecting adversarial examples (AEs) and then rejecting them (*i.e.*, detection/rejection mechanism) can improve the model's robustness against adversarial attacks. That is, the input will be rejected if the detector classifies it as an adversarial example. Popular detection techniques include Odds [60], which considers the difference between clean images and AEs in terms of log-odds; NIC [51], which checks channel invariants within deep neural networks (DNNs); GAT [88], which resorts to multiple binary classifiers; JTLA [58], which proposes a detection framework by employing internal layer representations, among others [15, 17, 34, 67, 85]. Unfortunately, existing detection methods are typically ineffective in defending against adaptive attacks. Besides, the detection/rejection mechanism cannot be generalized to domains where natural adversarial examples exist. Our work differs from previous solutions in two aspects. First, we introduce a novel Multi-head Self-Attention (MSA) mechanism by using the Guided Backpropagation technique, which can largely expose adversarial perturbations. Second, we incorporate the Soft-Nearest Neighbors (SNN) loss to maximize the differences between clean images and adversarial examples. These innovative designs enable our detector to effectively defend against adaptive attacks. Moreover, our newly developed adaptive ensemble further enhances our detector, empowering it to be applied to scenarios where rejecting input images is not allowed.

Adversarial Training Approaches. Adversarial training (AT) aims to improve the model's robustness against adversarial attacks by intentionally injecting adversarial examples into the training data. For example, PGD-AT [52] proposes a multi-step attack to find the worst case of training data, TRADES [90] addresses the limitation of PGD-AT by utilizing theoretically sound classification-calibrated loss, EAT [76] uses an ensemble of different DNNs to produce the threat model, FAT [82] reduces the computational overhead of AT by utilizing FGSM attack with the random initialization, LAS-AWP [29] boosts AT with a learnable attack strategy, Sub-AT [35] constrains AT in a well-designed subspace, and many others [1, 7, 16, 22, 23, 30, 42–44, 53, 66, 79, 83, 84, 93, 94]. However, prior ATs suffer from the dilemma of balancing the trade-off between standard accuracy and adversarial robustness. Besides, their improved robustness is vulnerable to adaptive attacks. In contrast, our work introduces a ViT architecture consisting of a detector and a classifier, connected by a newly developed adaptive ensemble, able to boost AT to defend against adaptive attacks. Meanwhile, it lowers the standard accuracy degradation by employing two encoders for extracting visual representations respectively from clean images and adversarial examples, empowering our ViT architecture to enjoy a better trade-off between accuracy and robustness.

3 PRELIMINARY: ONE-STEP LEAST-LIKELY ADVERSARIAL TRAINING

Adversarial training (AT) improves the model's robustness against adversarial attacks by feeding adversarial examples into the training set. Given a model f with parameters $\boldsymbol{\theta}$, a dataset with N samples, *i.e.*, $\mathbb{X} = \{(\boldsymbol{x}_i, y_i) \mid i \in \{1, 2, \ldots, N\}\}$, the cross-entropy loss function $\mathcal{L}$, and a threat model $\boldsymbol{\Delta}$, AT aims to solve the following inner-maximization problem and outer-minimization problem, *i.e.*,

$$\min_{\boldsymbol{\theta}} \sum_{i}^{N} \max_{\boldsymbol{\delta} \in \boldsymbol{\Delta}} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i + \boldsymbol{\delta}),\ y_i), \tag{1}$$

where the inner problem aims to find the worst-case training data for the given model, and the outer problem aims to improve the model's performance on such data. Recently, one-step Fast Adversarial Training (FAT) [82] is popular due to its computational efficiency. FAT sets the threat model under a small and l_∞ constraint ϵ, *i.e.*, $\boldsymbol{\Delta} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \leq \epsilon\}$, by performing Fast Gradient Sign Method (FGSM) [18] with the random initialization, *i.e.*,

$$\begin{aligned} \boldsymbol{\delta} &= \text{Uniform}(-\epsilon, \epsilon) + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}}\, \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i),\ y_i)), \\ \boldsymbol{\delta} &= \max(\min(\boldsymbol{\delta}, \epsilon), -\epsilon), \end{aligned} \tag{2}$$

where *Uniform* denotes the uniform distribution and *sign* is the sign function. Notably, the second row in Eq. (2) serves to project the perturbation $\boldsymbol{\delta}$ back into the l_∞ ball around the data $\boldsymbol{x}_i$.

To find the worst-case adversarial examples, we extend FAT by performing the least-likely targeted attacks, inspired by prior studies [33, 76]. That is, given an input $\boldsymbol{x}_i$, we perform targeted FGSM by setting the targeted label as its least-likely class, *i.e.*, $y_i^{ll} = \arg\min f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$, arriving at,

$$\begin{aligned} \boldsymbol{\delta} &= \text{Uniform}(-\epsilon, \epsilon) + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i),\ y_i^{ll})), \\ \boldsymbol{\delta} &= \max(\min(\boldsymbol{\delta}, \epsilon), -\epsilon), \end{aligned} \tag{3}$$

Our one-step least-likely adversarial training is to utilize Eq.(3) to produce the threat model.

4 OUR APPROACHES

4.1 Problem Statement

We consider a set of N samples, *i.e.*, $\mathbb{X} = \{(\boldsymbol{x}_i,\ y_i) \mid i \in \{1, 2, \ldots, N\}\}$, where $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C_H}$ is an input image with the resolution of (H, W) and the channel count of C_H, and $y \in [C]$ denotes its label. For notational convenience, we let $d = H \times W \times C_H$. A classifier is a function $f_{\boldsymbol{\theta}}: \mathbb{R}^d \to [C]$, parameterized by a neural network. We consider two types of inputs, *i.e.*, a clean image $\boldsymbol{x}^{\text{cln}}$ sampled from the standard distribution $\mathcal{D}_{\text{std}}$ and an adversarial example $\boldsymbol{x}^{\text{adv}}$ sampled from the adversarial distribution $\mathcal{D}_{\text{adv}}$. We assume $\mathcal{D}_{\text{std}}$ and $\mathcal{D}_{\text{adv}}$ follow different distributions. The clean image $\boldsymbol{x}^{\text{cln}}$ itself or its augmented variant can be the input, while the adversarial example $\boldsymbol{x}^{\text{adv}}$ is a malicious version of $\boldsymbol{x}$ within a small distance. That is, for some metric d, we have $d(\boldsymbol{x}, \boldsymbol{x}^{\text{adv}}) \leq \epsilon$, but $\boldsymbol{x}^{\text{adv}}$ can mislead conventional classifiers. Parameterized by another neural network, a detector $g_{\boldsymbol{\phi}}$ is to tell whether an input image is a clean image or not, *i.e.*, $g_{\boldsymbol{\phi}}: \mathbb{R}^d \to \{\pm 1\}$, where $+1$ and -1 indicate a clean image and an adversarial example, respectively. The binary indicator function $\mathbb{1}_{\{\cdot\}}$ is 1 if both the detector $g_{\boldsymbol{\phi}}$ and the classifier $f_{\boldsymbol{\theta}}$ make correct predictions. We follow previous studies [52, 90] by referring standard accuracy, and adversarial robustness, as classification accuracy on clean images and adversarial examples, respectively.

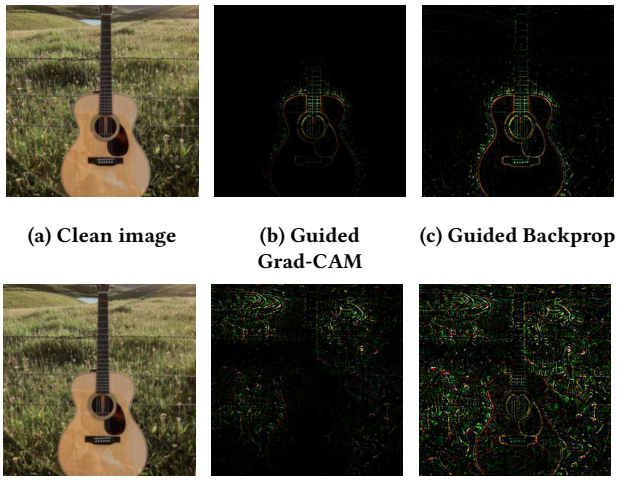
4.2 Detector

Parameterized by a neural network with parameters $\boldsymbol{\phi}$, the detector $g_{\boldsymbol{\phi}}: \mathbb{R}^d \to \{\pm 1\}$ is to determine whether the input is a clean image or not, where $+1$ and -1 respectively represent a clean image and an adversarial example, *i.e.*,

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}) = \begin{cases} +1, & \text{if } \boldsymbol{x} \text{ is a clean image} \\ -1, & \text{otherwise.} \end{cases} \tag{4}$$

Aiming to generalize the robust model to critical domains (*e.g.*, autonomous driving), the input will not be rejected in this work. Instead, we have modified it to output an estimated probability of $p \in [0, 1]$ for clean images and $1 - p$ for adversarial examples.

The design of our detector architecture is motivated by our empirical observation in that *the adversarial perturbation is detectable after Guided Backpropagation visualization*. Due to the small distance between a clean image and its corresponding adversarial example, their difference is notoriously imperceptible (see Figures 1a and 1d), making it theoretically hard to detect adversarial examples [74]. In our empirical study, we resort to Guided Backpropagation [71] to visualize the difference between a clean image and an adversarial example. Interestingly, we have discovered that after Guided Backpropagation visualization on the adversarial example, its adversarial perturbation is quite noticeable; see Figure 1c versus Figure 1f, *i.e.*, visualization on a clean image versus on its adversarial example. Notably, our experiments also include the visualization comparison of Guided Grad-CAM [65], developed recently; see Figure 1b versus Figure 1e. However, Guided Grad-CAM exhibits inferior performance (compared to Guided Backpropagation) in terms of exposing adversarial perturbation. This empirical study motivates us to maximize the difference between clean images and adversarial examples by using Guided Backpropagation visualization.



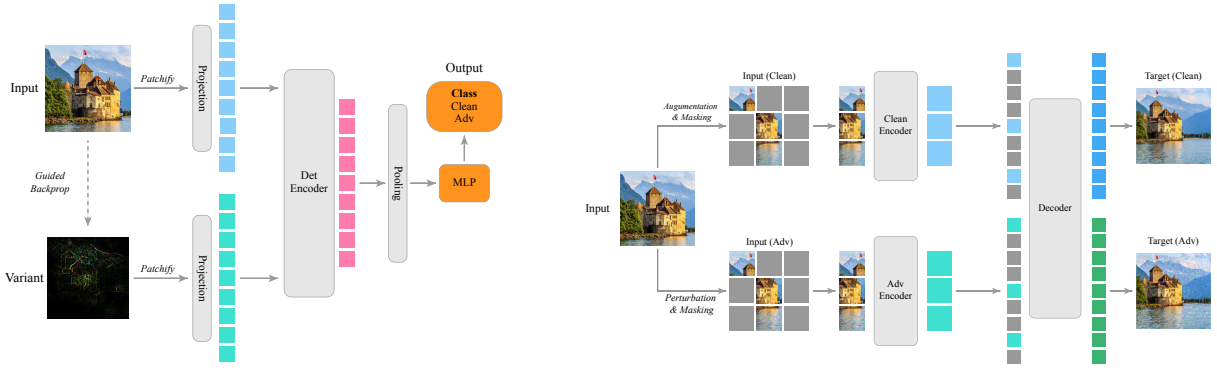
(a) Clean image (b) Guided Grad-CAM (c) Guided Backprop

(d) Adversarial example (e) Guided Grad-CAM (f) Guided Backprop

Figure 1: Visualizations on the clean image (Top) and the adversarial example (Bottom). Left: Original clean image and adversarial example. Middle: Guided Grad-CAM visualization. Right: Guided Backpropagation visualization.

Figure 2a illustrates our detector architecture. Given an input image $\boldsymbol{x} \in \mathbb{R}^d$, we perform Guided Backpropagation on the original image, arriving at an input variant $\boldsymbol{x}' \in \mathbb{R}^d$. Following the standard Vision Transformers (ViT) [12], we patchify the two inputs into two sets of image patches and embed them via linear projection, arriving at two sets of patch embeddings, *i.e.*, $\mathbf{E}_p \in \mathbb{R}^{M \times D}$ and $\mathbf{E}'_p \in \mathbb{R}^{M \times D}$, respectively for the original input and its input variant. Here, M represents the number of patches and D indicates the hidden dimension. Driven by the above empirical observation, a naive idea to expose adversarial perturbation is to add two sets of patch embeddings. However, our empirical results show that this simple solution cannot achieve satisfactory performance. To address this issue, we propose a novel Multi-head Self-Attention (MSA) [78] to consider two sets of patch embeddings simultaneously, inspired by recent studies [4, 46, 57]. Let $\mathbf{E} = \mathbf{E}_p + \mathbf{E}_{\text{pos}}$ and $\mathbf{E}' = \mathbf{E}'_p + \mathbf{E}_{\text{pos}}$ respectively represent two sets of patch embeddings after adding positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{M \times D}$, our proposed MSA can be expressed as follows:

$$\begin{aligned} \text{MSA}\ (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T + \boldsymbol{B}}{\sqrt{d}})\boldsymbol{V}, \\ \boldsymbol{Q} = \boldsymbol{W}_Q \cdot \boldsymbol{E},\ \boldsymbol{K} &= \boldsymbol{W}_K \cdot \boldsymbol{E},\ \boldsymbol{V} = \boldsymbol{W}_V \cdot \boldsymbol{E}. \end{aligned} \tag{5}$$



(a) Detector **(b) Classifer**

Figure 2: Our model architecture: (a) detector and (b) classifier during the pre-training stage.

Here, $\boldsymbol{B} = \boldsymbol{W}_B \cdot \boldsymbol{E}'$ is the relative detection bias obtained from the Guided Backpropagation-based input variant. $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$, and $\boldsymbol{W}_B$ are learnable projection matrices, similar to those in prior studies [59, 78]. The intuition underlying Eq. (5) is that we aim to expose adversarial perturbation by adding the relative bias obtained from Guided Backpropagation visualization. After encoding, we follow Masked Autoencoders (MAE) [20] by performing global average pooling on the full set of encoded patch embeddings, with the resulting token fed into an MLP (*i.e.*, multiple-layer perceptron) for telling whether the input is a clean image or not.

Aiming to further differentiate adversarial examples from clean images, we propose a novel loss function to train our detector, including a Cross-Entropy (CE) Loss $\mathcal{L}_{\text{ce}}$ and a Soft-Nearest Neighbors (SNN) loss $\mathcal{L}_{\text{snn}}$ [14, 64], for jointly penalizing the detection error and the similarity level between the clean image and the adversarial example, *i.e.*,

$$\mathcal{L}_{\text{det}} = (1-\lambda)\cdot\mathcal{L}_{\text{ce}}(g_{\boldsymbol{\phi}}(\boldsymbol{x}), y^{\text{det}}) + \lambda\cdot\mathcal{L}_{\text{snn}}(z^{\text{cln}}, z^{\text{adv}}), \quad (6)$$

where $\lambda \in (0,1)$ is a hyperparameter to control the penalty degree of the two terms, and z^{cln} and z^{adv} denote the global representations, *i.e.*, the global average pooling of encoded representations, for clean images and adversarial examples, respectively.

The SNN loss is a variant of contrastive loss, allowing for the inclusion of multiple positive pairs. We regard members belonging to the same determined class (*e.g.*, two clean images) as positive pairs, while members belonging to different determined classes (*e.g.*, a clean image and an adversarial example) as negative pairs. Given a mini-batch of $2B$ samples, with one half being clean images, *i.e.*, $\{(\boldsymbol{x}_i,\ y_i^{\text{det}}{=}1)\}_{i=1}^{B}$, and the other half of adversarial examples, *i.e.*, $\{(\boldsymbol{x}_i^{\text{adv}},\ y_i^{\text{det}}{=}{-1})\}_{i=B+1}^{2B}$, the SNN loss at temperature τ is defined below:

$$\mathcal{L}_{\text{snn}} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\sum_{j=1,i\neq j,y_i^{\text{det}}=y_j^{\text{det}}}^{2B}\exp(-\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_j)/\tau)}{\sum_{j=1,i\neq k}^{2B}\exp(-\text{sim}(\boldsymbol{z}_i,\boldsymbol{z}_k)/\tau)}, \quad (7)$$

where $\boldsymbol{z}_i$ is the visual representations for the input $\boldsymbol{x}_i$ and the similarity metric $\text{sim}(\cdot,\cdot)$ is measured by the cosine distance. The SNN loss enforces each point to be closer to its positive pairs than to its negative pairs. In other words, the SNN loss penalizes the similarity level between clean images and adversarial examples, making adversarial examples more discernible by our detector.

4.3 Classifer

Inspired by self-supervised learning for vision tasks [3, 8, 20], we separate our adversarial training into two stages, *i.e.*, pre-training and fine-tuning, for learning high-quality visual representations and fine-tuning a robust classifier, respectively.

Pre-training. Our classifier architecture for the pre-training is inspired by MAE [20]. Different from MAE, we utilize two encoders, denoted as the clean encoder and the adversarial encoder, for learning visual representations from clean images and adversarial examples, respectively. The decoder aims to reconstruct the original inputs from the visual representations encoded by the two encoders. Figure 2b shows the classifier architecture during the pre-training. Given an input image $\boldsymbol{x} \in \mathbb{R}^d$, let $\boldsymbol{x}^{\text{cln}}$ and $\boldsymbol{x}^{\text{adv}}$ denote its clean and adversarial variants, respectively, with the clean variant obtained by augmenting the original input. Regarding the clean variant $\boldsymbol{x}^{\text{cln}}$, we randomly mask out a large proportion of image patches (*e.g.*, 75%) and then feed the subset of visible patches into the clean encoder. The masked tokens are inserted into corresponding positions after the encoder. Finally, the decoder reconstructs the clean variant $\bar{\boldsymbol{x}}^{\text{cln}}$ from the full set of image patches, including encoded visible patches and masked tokens. The reconstruction of the adversarial variant $\boldsymbol{x}^{\text{adv}}$ follows a similar procedure, except that its visible patches are encoded by the adversarial encoder. Notably, the position of masked image patches in the adversarial variant $\boldsymbol{x}^{\text{adv}}$ is the same as that in the clean variant $\boldsymbol{x}^{\text{cln}}$ in order to minimize their visual representation difference during the pre-training.

Let $\bar{\boldsymbol{z}}^{\text{cln}}$ and $\bar{\boldsymbol{z}}^{\text{adv}}$ respectively denote the global representations of clean and adversarial variants, obtained by performing global average pooling on the decoder's input sequence. Our design utilizes a new loss function to learn visual representations by simultaneously minimizing the reconstruction error and the visual representation difference, *i.e.*,

$$\mathcal{L}_{\text{enc}} = (1-\Omega)\cdot\mathcal{L}_{\text{rec}}(\boldsymbol{x},\bar{\boldsymbol{x}}) + \Omega\cdot\mathcal{L}_{\text{cl}}(\bar{\boldsymbol{z}}^{\text{cln}},\bar{\boldsymbol{z}}^{\text{adv}}), \quad (8)$$

where $\Omega \in (0,1)$ is a hyperparameter and $\bar{\boldsymbol{x}}$ is the reconstructed image. $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{cl}}$ denote the reconstruction loss and the contrastive loss, respectively. Given a set of B input images, we first generate their adversarial variants, arriving at a mini-batch of $2B$ samples, consisting of B clean variants $\{\boldsymbol{x}_i^{\text{cln}}\}_{i=1}^{B}$ and B adversarial variants $\{\boldsymbol{x}_i^{\text{adv}}\}_{i=B+1}^{2B}$. We consider the form of contrastive loss in

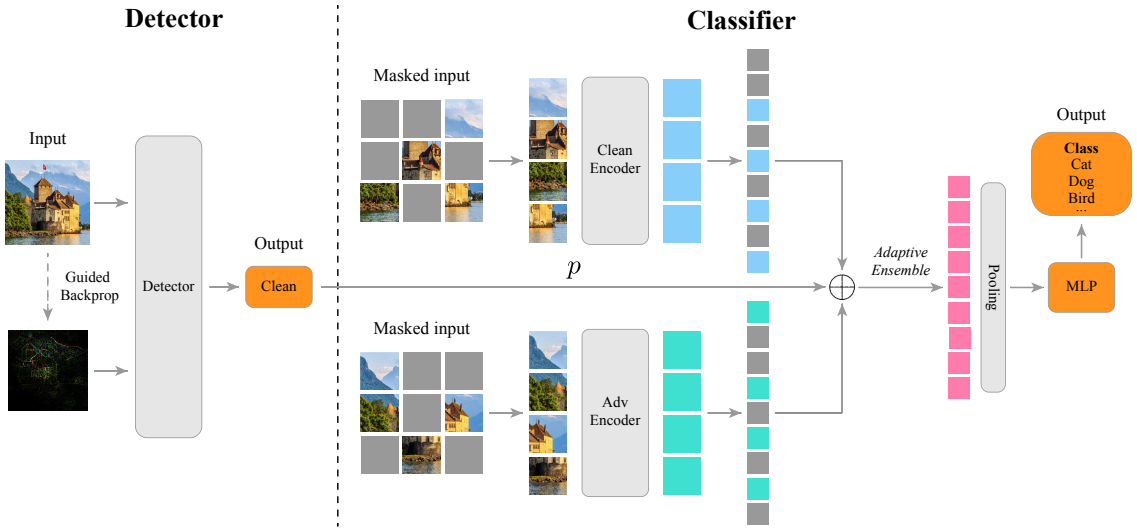


Figure 3: Illustration of our model architecture during the fine-tuning stage.

SimCLR [8], and define our contrastive loss at temperature τ as follows:

$$\ell(i,j) = -\log \frac{\exp(\text{sim}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j)/\tau)}{\sum_{i\neq k, k=1,\ldots,2B} \exp(\text{sim}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_k)/\tau)},$$
$$\mathcal{L}_{\text{cl}} = \frac{1}{2B}\sum_{k=1}^{B} \left[\ell(k, k+B) + \ell(k+B, k)\right], \tag{9}$$

where $\bar{z}_i$ denotes visual representations for $\boldsymbol{x}_i^{\text{cln}}$ (or $\boldsymbol{x}_i^{\text{adv}}$) and the similarity level $\text{sim}(\cdot,\cdot)$ is measured by the cosine distance. In particular, we regard clean and adversarial variants from the same input as positive pairs, while the rest in the same batch are negative pairs. Hence, the loss value decreases when visual representations for the clean and the adversarial variants of the same input become more similar.

Fine-tuning. The detector and the classifier (including two encoders and one decoder) are trained jointly in the pre-training stage. After that, we drop the decoder and freeze the weights in the well-trained detector and two encoders, with Figure 3 depicting our model architecture during the fine-tuning stage. Different from MAE, which encodes the full set of image patches during the fine-tuning, our approach randomly masks out a relatively small proportion of image patches (*e.g.*, 45%), aiming to eliminate the potential adversarial effect if the input is an adversarial example.

Given an input image $(\boldsymbol{x}, y^{\text{cls}})$, where $\boldsymbol{x} \in \mathbb{R}^d$ is either a clean image or an adversarial example with the label $y^{\text{cls}} \in [C]$, we randomly mask the input image twice, arriving at two different masked inputs. Two subsets of visible patches from the two masked inputs are fed into the clean and the adversarial encoders, respectively. The masked tokens are introduced onto their corresponding positions after the encoder, obtaining two full sets of visual representations, *i.e.*, $\hat{\boldsymbol{z}}^{\text{cln}}$ and $\hat{\boldsymbol{z}}^{\text{adv}}$ which are partially encoded by the clean and the adversarial encoders, respectively. We then perform the global average pooling on the *adaptive ensemble* of $\hat{\boldsymbol{z}}^{\text{cln}}$ and $\hat{\boldsymbol{z}}^{\text{adv}}$, with the result fed into an MLP for classification.

Adaptive Ensemble. Although randomly masking an input image can eliminate the potential adversarial effect, this way inevitably hurts standard accuracy during the fine-tuning. In this paper, we propose *adaptive ensemble* [37] to tackle this issue. That is, the global representation for an input image is derived from the sum of $\hat{\boldsymbol{z}}^{\text{cln}}$ and $\hat{\boldsymbol{z}}^{\text{adv}}$ with an adaptive factor $p \in [0,1]$, where $\hat{\boldsymbol{z}}^{\text{cln}}$ and $\hat{\boldsymbol{z}}^{\text{adv}}$ are visual representations encoded by the clean and the adversarial encoders, respectively, and p is the probability of the input image being a clean image estimated by our detector.

Let A be a full set of image patches and V be a subset of A, including visible patches only. $\mathbb{1}_V(\cdot)$ is the indicator function for evaluating whether an image patch is visible. Hence, for each image patch of A, we have,

$$\mathbb{1}_V(i) = \begin{cases} 1, & \text{if the patch is visible} \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, 2, \ldots, M, \tag{10}$$

where M is the number of image patches, *i.e.*, $|A|$. For notational convenience, we let $\mathbb{1}_V^{\text{cln}}$ indicate visible patches fed into the clean encoder. Likewise, $\mathbb{1}_V^{\text{adv}}$ indicates visible patches fed into the adversarial encoder. Let $\hat{\boldsymbol{z}}_i$ be the visual representation of the i-th image patch, with $i \in \{1, 2, \ldots, M\}$. Our adaptive ensemble is defined by:

$$\hat{\boldsymbol{z}}_i = \frac{p \cdot \mathbb{1}_V^{\text{cln}}(i) \cdot \hat{\boldsymbol{z}}_i^{\text{cln}} + (1-p) \cdot \mathbb{1}_V^{\text{adv}}(i) \cdot \hat{\boldsymbol{z}}_i^{\text{adv}}}{\max\left(p \cdot \mathbb{1}_V^{\text{cln}}(i) + (1-p) \cdot \mathbb{1}_V^{\text{adv}}(i),\ \epsilon\right)}, \tag{11}$$

where the denominator serves to normalize the adaptive ensemble of $\hat{\boldsymbol{z}}_i^{\text{cln}}$ and $\hat{\boldsymbol{z}}_i^{\text{adv}}$, and ϵ is a small value to avoid divison by zero (*i.e.*, $\epsilon = 1e-12$ in this paper). The intuition underlying Eq. (11) is that if our detector has a high confidence that the input is a clean image (*i.e.*, p is large), the global representation $\hat{\boldsymbol{z}}_i$ will be mostly encoded by the clean encoder. Otherwise, $\hat{\boldsymbol{z}}_i$ will be mainly encoded by the adversarial encoder. In addition, as our pre-training encourages the similarity level of the clean and the adversarial variants from a given input (see Eq. (8) and Eq. (9)), and two different masked inputs exist upon the fine-tuning, the invisible image patches in one masked input can be glimpsed from the other masked input.

5 EXPERIMENTS AND RESULTS

5.1 Experimental Setup

Datasets. We conduct experiments on three widely-used benchmarks. (i) **CIFAR-10** [31]: 60,000 32x32 RGB images of 10 classes. (ii) **CIFAR-100** [31]: 60,000 32x32 RGB examples in 100 categories. (iii) **Tiny-ImageNet** [10]: 120,000 64x64 RGB images of 200 classes.

Compared Methods. We compare our approach with four detection methods, *i.e.*, **Odds** [60], **NIC** [51], **GAT** [88], and **JTLA** [58]. We compare our approach with five adversarial training (AT) counterparts: **PGD-AT** [52], **TRADES** [90], **FAT** [82], **Sub-AT** [35], and **LAS-AWP** [29], to exhibit how it boosts the ViT's robustness.

Evaluation. We consider three state-of-the-art adaptive attacks, *i.e.*, **AutoAttack** [9], **Adaptive Auto Attack (A^3)** [87], and **Parameter-Free Adaptive Auto Attack (PF-A^3)** [45], for evaluating detection accuracy and adversarial robustness. The attack constraint, if not specified, is set to $\epsilon = 8/255$.

Model Size. We build our detector and classifier on top of Vision Transformers (ViT), with their architectures following ViT [12] and MAE [20], respectively. Our model size is pruned down to as small as possible in order to conduct a fair comparison with baselines. Table 1 lists the model size details. Our architecture consists of a detector and a classifier (including two encoders and one decoder), with $54.0M$ parameters in total. To conduct a fair comparison, existing adversarial training baselines use the ViT-Base model [12] with total parameters of $85.6M$ as the backbone network.

Hyperparameters. For all our models, if not specified, we use AdamW [48] with β_1=0.9, β_2=0.999, the weight decay of 0.05, and a batch size of 512. During the pre-training, the detector and the classifier (*i.e.*, two encoders and one decoder) are trained jointly. For the detector, we follow the setting in [19] by setting the epochs of 100, the base learning rate of $1e-3$, the linear warmup epochs of 5, and the cosine decay schedule [47]. For the classifier, by contrast, we pre-train it for 200 epochs, with the base learning rate of $1e-4$, the linear warmup of 20 epochs, and a masking ratio of 75%. After pre-training, we drop the decoder and freeze the weights on the detector and the two encoders. Then, we finetune the classifier for 100 epochs, with the base learning rate of $1e-3$, the linear warmup of 5, and the cosine decay schedule, and a masking ratio of 45%. The patch size is set to 4 (or 8) for CIFAR-10/CIFAR-100 (or Tiny-ImageNet). We grid-search hyperparamters λ in Eq. (6) and Ω in Eq. (8) of Section 4 and empirically set λ to 0.15 and Ω to 0.35 for all datasets.

5.2 Overall Performance on Our Classifier

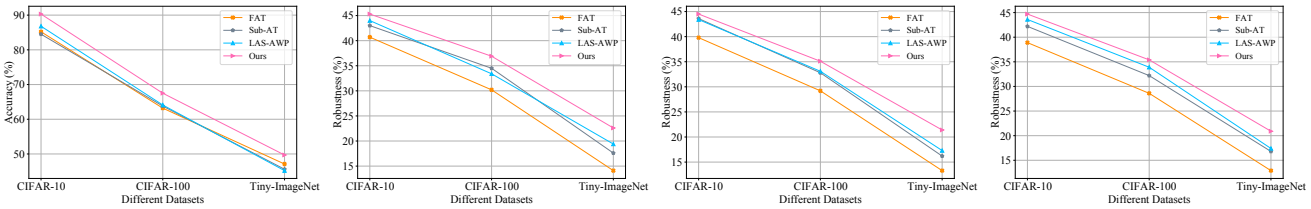
Overall Comparisons on CIFAR-10. We first conduct extensive experiments on CIFAR-10 and compare our approach to its state-of-the-art adversarial training (AT) counterparts listed in Section 5.1 in terms of standard accuracy and adversarial robustness under attack constraints of $\epsilon = 4/255$ and of $\epsilon = 8/255$. Table 2 lists comparative results. It is observed that our approach achieves the best performance under all three scenarios. In particular, our approach achieves the standard accuracy of 90.3%, outperforming the best competitor (*i.e.*, LAS-AWP) by 3.5%. This is contributed by employing two encoders to extract visual representations respectively from clean images and adversarial examples, able to significantly mitigate the adverse effect of adversarial training on standard accuracy. Besides, when the attack constraint is set to $\epsilon = 4/255$, our approach achieves the best robustness of 49.8%, 49.5%, and 48.1% against AutoAttack, Adaptive Auto Attack (A^3), and Parameter-Free Adaptive Auto Attack (PF-A^3), respectively. Our method significantly surpasses all its counterparts. For example, it outperforms two recent state-of-the-arts, *i.e.*, Sub-AT and LAS-AWP, respectively by 2.6% and 4.5% under the attack of A^3. Thirdly, increasing the attack constraint to $\epsilon = 8/255$ results in the decrease of adversarial robustness. But our approach still maintains the best robustness of 45.3%, 44.5%, and 44.7% under the attack of AutoAttack, A^3, and PF-A^3, respectively. The comparative results demonstrate that our masked adaptive ensemble is robust enough to withstand strong white-box attacks. This is because masking a small proportion of image patches can significantly mitigate the adversarial effect of malicious inputs.

Overall Comparisons on CIFAR-100 and Tiny-ImageNet. Here, we conduct a comprehensive comparison between our approach and adversarial training (AT) counterparts on CIFAR-100 and Tiny-ImageNet datasets. Table 3 lists the comparative results. On CIFAR-100, we observed that our approach achieves the best standard accuracy of 67.5%, outperforming the best competitor (*i.e.*, LAS-AWP) by 3.4%. Meanwhile, our method achieves the best robustness of 36.9%, 35.1%, and 35.4% under the attack of AutoAttack, A^3, and PF-A^3, respectively. This confirms that our approach can achieve a decent standard accuracy and robustness when being generalized to the dataset with large classes. On the Tiny-ImageNet dataset, both our approach and the baseline methods experience a decrease in performance. However, our proposed method still achieves the highest standard accuracy of 49.7%, which outperforms the best baseline (*i.e.*, FAT), by 2.6%. Moreover, all baselines suffer from a poor robustness on the Tiny-ImageNet dataset (*i.e.*, $\leq$ 20.0%), while our approach maintains a decent robustness of 22.6%, 21.4%, and 20.9% under the attack of AutoAttack, A^3, and PF-A^3, respectively.

Performance Stability. We next conduct experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet to evaluate the performance stability under different scales of datasets and different types of adaptive attacks. We compare our approach with three baselines, *i.e.*, FAT, Sub-AT, and LAS-AWP. Figures 4a, 4b, 4c and 4d illustrate the comparative results of standard accuracy, as well as robustness against AutoAttack, A^3, and PF-A^3, respectively. We have three discoveries. First, as depicted in Figure 4a, our approach (*i.e.*, the pink line) achieves the best standard accuracies of 90.3%, 67.5%, and 49.7% under CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. The empirical evidence verifies that our approach can maintain superior standard accuracy when generalized to large datasets. Second, on all three datasets, our approach achieves the best robustness under all adaptive attacks, as shown in Figures 4b, 4c and 4d. Take the robustness results under PF-A^3 (*i.e.*, Figure 4d) for example, our proposed masked adaptive ensemble achieves the robustness of 44.7%, 35.4%, and 20.9% on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively. These results outperform those of LAS-AWP (*i.e.*, the blue line), which is the best baseline, by 3.1%, 4.1%, and 3.5%, respectively. Third, when scaling up the dataset from CIFAR-10 to Tiny-ImageNet, our approach suffers from the least robustness degradation of 22.7%, 23.1%, and 23.8% under the

Table 1: Model Details used in our design

Model		Layer	Hidden Size	Head	MLP Size	Parameters
Detector	ViT-Tiny	12	192	3	768	7.8M
Classifier	Clean Encoder	12	384	3	1536	21.3M
	Adv Encoder	12	384	3	1536	21.3M
	Decoder	8	192	4	768	3.6M



(a) Standard Accuracy **(b) Under AutoAttack** **(c) Under A^3 attack** **(d) Under PF-A^3 attack**

Figure 4: Performance stability under different datasets and different adaptive attacks.

Table 2: Overall comparative results of standard accuracy and adversarial robustness on CIFAR-10 under attack constraints of $\epsilon = 4/255$ and of $\epsilon = 8/255$, with best results shown in bold

Method	Standard Accuracy	Robustness ($\epsilon = 4/255$)			Robustness ($\epsilon = 8/255$)		
		AutoAttack	A^3	PF-A^3	AutoAttack	A^3	PF-A^3
PGD-AT	83.7	45.1	43.4	43.5	41.5	40.3	40.9
TRADES	84.9	46.6	43.5	43.6	41.5	40.7	40.3
FAT	85.2	43.8	42.1	43.4	40.7	39.8	38.9
Sub-AT	84.5	47.4	46.9	45.1	43.3	43.6	42.2
LAS-AWP	86.8	46.9	45.0	47.7	44.0	43.4	43.6
Ours	**90.3**	**49.8**	**49.5**	**48.1**	**45.3**	**44.5**	**44.7**

Table 3: Overall comparisons on CIFAR-100 and Tiny-ImageNet, with best results shown in bold

Method	CIFAR-100				Tiny-ImageNet			
	Standard Accuracy	AutoAttack	A^3	PF-A^3	Standard Accuracy	AutoAttack	A^3	PF-A^3
PGD-AT	62.5	31.9	31.5	31.1	42.9	17.2	16.4	16.8
TRADES	61.8	32.5	31.3	31.6	44.1	15.2	14.7	14.1
FAT	63.2	30.2	29.2	28.6	47.1	14.1	13.3	12.9
Sub-AT	63.8	34.5	32.8	32.2	45.6	17.6	16.2	16.8
LAS-AWP	64.1	33.4	33.1	33.9	45.2	19.4	17.3	17.4
Ours	**67.5**	**36.9**	**35.1**	**35.4**	**49.7**	**22.6**	**21.4**	**20.9**

attack of AutoAttack, A^3, and PF-A^3, respectively. These results confirm that in terms of robustness, our approach enjoys the best performance stability upon scaling up to large datasets.

5.3 Ablation Studies on Our Classifier

Pre-training: Contrastive Loss. We qualitatively and quantitatively exhibit the impact of our proposed loss, *i.e.*, Eq. (9), on learning visual representations. We first present the qualitative evaluations. Specifically, we reconstruct masked adversarial examples and compare reconstruction quality by utilizing our approach with/without the contrastive loss (CL) in SimCLR [8]. Figure 5 illustrates the qualitative results. For images on each row, from left to right, are original adversarial example, the masked input, the image generated by our approach without the CL (*i.e.*, w/o CL), and the image reconstructed by our approach with the CL (*i.e.*, w/ CL). We observed that when using the CL, our approach always achieves a better reconstruction quality; See the 3rd (and 7th) column versus the 4th (and 8th) column. Besides, we discovered that our approach (w/o CL), in some cases, reconstructs adversarial examples with poor quality; See the 3rd and 7th columns in the last row. By contrast, our method (w/ CL) still achieves a high reconstruction quality on these examples; See the 4th and 8th columns in the last row. These results demonstrate that our proposed loss can boost the performance when learning visual representations from adversarial examples.

Table 4: Ablation studies on the classifier, including (a) the contrastive loss (CL) in pre-training stage and (b) the adaptive ensemble (AE) in fine-tuning stage

Method	Standard Accuracy	Robustness	
		A^3	PF-A^3
w/o CL	74.6	35.7	34.2
w/ CL	79.5	41.8	42.6

(a) Pre-training

Method	Standard Accuracy	Robustness	
		A^3	PF-A^3
w/o AE	81.9	38.9	39.4
w/ AE	90.3	44.5	44.7

(b) Fine-tuning

Next, we conduct experiments on CIFAR-10 for quantitatively evaluating visual representations by using the linear probing accuracy. Specifically, we consider the standard accuracy as well as the robustness under the attack of A^3 and PF-A^3. Table 4a presents the experimental results. We observed that by utilizing the contrastive loss, our approach achieves performance improvement of 4.9%, 6.1%, and 8.4% on standard accuracy, robustness against A^3, and robustness against PF-A^3, respectively. These empirical results demonstrate the necessity and importance of our proposed loss for learning high-quality visual representations.

Fine-tuning: Adaptive Ensemble. Here, we conduct experiments to show the impact of our adaptive ensemble on the standard accuracy and the robustness. Table 4b lists the experimental results with/without our adaptive ensemble. Note that we employ the naive average ensemble when conducting experiments without

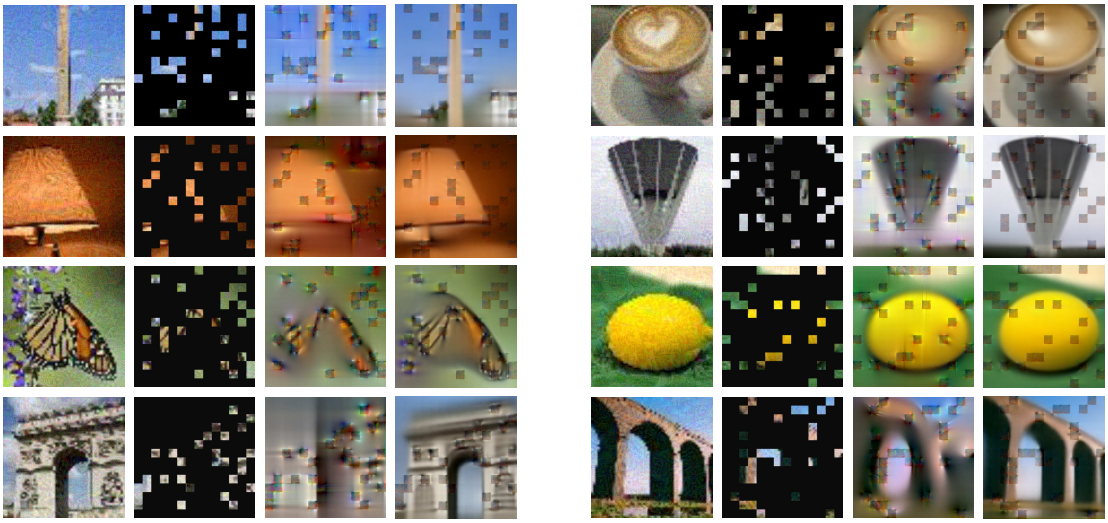


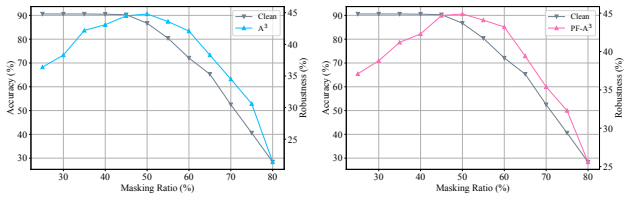
Figure 5: Comparison of the reconstruction quality from masked adversarial examples by employing our approach with/without the contrastive loss, denoted as our approach (w/ CL) and our approach (w/o CL), respectively. From left to right are the original adversarial example, the masked input, reconstruction by our approach (w/o CL), and reconstruction by our approach (w/ CL), respectively.

our adaptive ensemble. From Table 4b, we observed that our adaptive ensemble significantly benefits the standard accuracy, with 8.4% performance improvement. Meanwhile, it boosts adversarial robustness against A^3 by 5.6% and against PF-A^3 by 5.3%. This is because the adaptive factor p estimated by our detector can adaptively adjust the proportion of visual representations from clean and adversarial encoders, thereby significantly boosting the classification performance.

Fine-tuning: Masking Ratio. We conduct experiments on CIFAR-10 to explore how different masking ratios affect the performance of our approach during the finetuning. 12 groups of masking ratios are taken into account, ranging from 25% to 80%. Note that in the pre-training, we directly set the masking ratio to 75% by following MAE [20]; hence, no similar ablation study is required. Here, we consider the trade-off between standard accuracy and robustness (under A^3 and PF-A^3 attacks).

Figures 6a and 6b illustrate experimental results. In Figure 6a, we observed that increasing the masking ratio negatively affects standard accuracy (*i.e.*, the grey line) in all scenarios. In contrast, when the masking ratio is small (*i.e.*, $\leq 50\%$), a larger masking ratio benefits robustness against A^3 (*i.e.*, the blue line). But when the masking ratio is greater than 50%, increasing the masking ratio hurts this robustness. This is because a small subset of masked patches can eliminate the adversarial effect of adversarial attacks, while a large subset of masked patches would prevent our classifier from accurate classification. Clearly, our approach achieves the best trade-off on the masking ratio of 45%, with standard accuracy of 90.3% and robustness against A^3 of 44.5%.

Similarly, Figure 6b depicts robustness against PF-A^3 (*i.e.*, the pink line) under different masking ratios. We also include standard accuracy (similar to Figure 6a) for a better illustration of the trade-off. Obviously, when the masking ratio equals 45%, our approach



(a) Under A^3 attack **(b) Under PF-A^3 attack**

Figure 6: Illustration of how different masking ratios in the finetuning affect the performance.

achieves the best trade-off, with standard accuracy of 90.3% and robustness of 44.7%. Based on the above discussion, we set our masking ratio to 45% to ensure the best trade-off between standard accuracy and robustness (under adaptive attacks).

5.4 Evaluating Our Detector

In this section, we conduct experiments on CIFAR-10 for comparing our detector with four detection baselines, *i.e.*, **Odds** [60], **NIC** [51], **GAT** [88], and **JTLA** [58]. Three aforementioned adaptive attacks under two small attack constraints, *i.e.*, $\epsilon = 2/255$ and $\epsilon = 4/255$, are used for evaluating detection accuracy. Table 5 lists the detection accuracy values under different attack methods. We observed that our detector achieves the best detection accuracy under all scenarios. Specifically, our approach achieves the best detection accuracy of 99.4% under the attack constraint of $\epsilon = 4/255$ (see the 5th column). Decreasing the attack constraint to $2/255$ increases the detection difficulty, with our approach still maintaining the superior detection accuracy of 95.8% (see the 4th column) in the worst case. Besides, our detector outperforms all baselines, with the detection accuracy improvements ranging from 1.8% (*i.e.*, 95.9% vs. 94.1%, see the 3rd column) to 6.3% (*i.e.*, 96.4% vs. 90.1%, see the

Table 5: Comparisons of detection accuracy on CIFAR-10 under different adaptive attacks with best results shown in bold

Method	Attack Constraint ($\epsilon = 2/255$)			Attack Constraint ($\epsilon = 4/255$)		
	AutoAttack	A^3	PF-A^3	AutoAttack	A^3	PF-A^3
Odds	90.1	90.6	91.2	94.8	94.3	94.9
NIC	93.1	92.5	94.4	95.8	95.6	96.4
GAT	92.6	93.8	93.0	96.0	95.8	95.6
JTLA	94.3	94.1	93.9	95.6	96.4	96.2
Ours	**96.4**	**95.9**	**95.8**	**99.4**	**98.7**	**98.9**

Table 6: Ablation studies on our detector

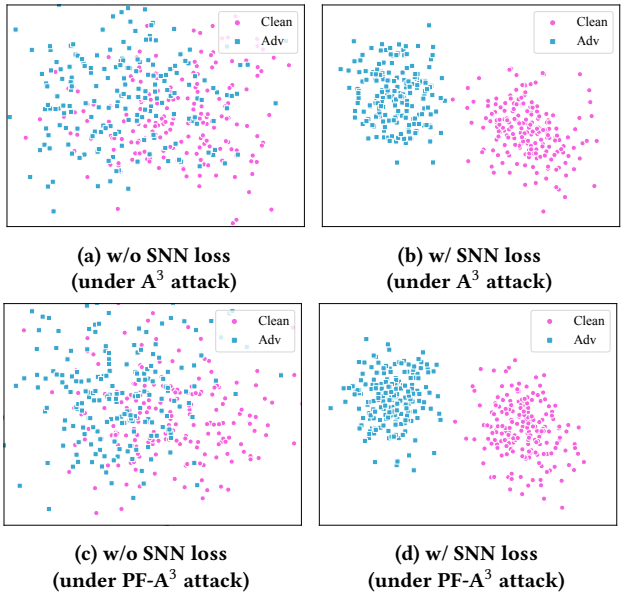
Method	A^3	PF-A^3
w/o GB	92.3	91.9
w/o MSA	92.9	92.8
Ours	98.7	98.9

2nd column). The statistical evidence exhibits that our two new designs for the detector, *i.e.*, the new Multi-head Self-Attention (MSA) mechanism and the proposed loss function, are effective for exposing adversarial perturbation, rendering our detector to better defend against adaptive attacks.

5.5 Ablation Studies on Our Detector

MSA on the Detection Accuracy. Here, we empirically show how our developed MSA mechanism affects the detection accuracy under the attack of A^3 and PF-A^3. We consider two scenarios. First, we remove the Guided Backpropagation (GB) variant to validate whether it benefits the detection of adversarial examples, denoted as "w/o GB". Second, we discard our proposed MSA and instead naively add two sets of patch embeddings respectively from the clean image and the GB variant, denoted as "w/o MSA". Table 6 lists the experimental results. We discovered that simply adding two sets of patch embeddings only marginally improves the detection accuracy of 0.6% (or 0.9%) under the A^3 (or PF-A^3) attack (see "w/o GB" vs. "w/o MSA"). Equipped with our MSA mechanism, in sharp contrast, the Guided Backpropagation technique can significantly benefit the detection task, with the detection accuracy improvement of 6.4% (or 7.0%) under the A^3 (or PF-A^3) attack (see "w/o GB" vs. "Ours"). These results confirm that (i) the Guided Backpropagation technique can help expose adversarial perturbation and (ii) our proposed MSA can significantly boost the detector's robustness against adaptive attacks.

SNN Loss on Visual Representations. Here, we reveal the effect of our proposed loss, *i.e.*, Eq. (6), on detecting adversarial examples. We consider how our detector with or without the Soft-Nearest Neighbors (SNN) loss affects the resulting representation space. In particular, we employ t-SNE visualization [77] on 200 clean images randomly sampled from CIFAR-10 and 200 adversarial examples generated either by the A^3 attack or by the PF-A^3 attack. Figures 7a and 7b depict the results by using the A^3 attack, while Figures 7c and 7d present the results by employing the PF-A^3 attack. We observed that without the SNN loss, the representations for clean images and adversarial examples are highly entangled; see Figures 7a and 7c. In sharp contrast, by minimizing the SNN loss, the representations for clean images and adversarial examples are mutually isolated, as shown in Figures 7b and 7d, making adversarial examples detectable.



(a) w/o SNN loss (under A^3 attack) **(b) w/ SNN loss (under A^3 attack)**

(c) w/o SNN loss (under PF-A^3 attack) **(d) w/ SNN loss (under PF-A^3 attack)**

Figure 7: t-SNE visualization on CIFAR-10 by using our detector with/without SNN loss. For each experiment, we perform t-SNE visualization on 200 clean images and 200 adversarial examples generated either by the A^3 attack, *i.e.*, (a) and (b), or by the PF-A^3 attack, *i.e.*, (c) and (d).

6 CONCLUSION

This article has proposed a novel Vison Transformers (ViT) architecture, including a detector and a classifier, which are bridged by a newly developed adaptive ensemble. This ViT architecture enables us to boost adversarial training to defend against adaptive attacks, and to achieve a better trade-off between standard accuracy and robustness. Our key idea includes introducing a novel Multi-head Self-Attention (MSA) mechanism to expose adversarial perturbations for better detection and employing two decoders to extract visual representations respectively from clean images and adversarial examples so as to reduce the negative effect of adversarial training on standard accuracy. Meanwhile, our adaptive ensemble lowers potential adversarial effects upon encountering adversarial examples by masking out a random subset of image patches across input data. Extensive experiments have been conducted for evaluation, showing that our solutions significantly outperform their state-of-the-art counterparts in terms of standard accuracy and robustness.

ACKNOWLEDGMENTS

This work was supported in part by NSF under Grants 2019511, 2348452, and 2315613. Any opinions and findings expressed in the paper are those of the authors and do not necessarily reflect the views of funding agencies.

REFERENCES

[1] Maksym Andriushchenko and Nicolas Flammarion. 2020. Understanding and Improving Fast Adversarial Training. In *NeurIPS*.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In *ICCV*.

[3] Hangbo Bao, Li Dong, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *ICLR*.

[4] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-Masked Language Models for Unified Language Model Pre-Training. In *ICML*.

[5] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. In *Symposium on Security and Privacy (SP)*.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *International Conference on Computer Vision (ICCV)*.

[7] Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. 2020. More data can expand the generalization gap between adversarially robust and standard models. In *ICML*.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*.

[9] Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

[11] Xiaoyi Dong, Dongdong Chen, Jianmin Bao, Chuan Qin, Lu Yuan, Weiming Zhang, Nenghai Yu, and Dong Chen. 2020. GreedyFool: Distortion-Aware Sparse Adversarial Attack. In *NeurIPS*.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale Vision Transformers. In *International Conference on Computer Vision (ICCV)*.

[14] Nicholas Frosst, Nicolas Papernot, and Geoffrey E. Hinton. 2019. Analyzing and Improving Representations with the Soft Nearest Neighbor Loss. In *ICML*.

[15] Hao Fu, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. 2023. Differential analysis of triggers and benign features for black-box DNN backdoor detection. *IEEE Transactions on Information Forensics and Security* (2023).

[16] Hao Fu, Alireza Sarmadi, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. 2023. Mitigating Backdoor Attacks on Deep Neural Networks. In *Embedded Machine Learning for Cyber-Physical, IoT, and Edge Computing: Use Cases and Emerging Challenges*. 395–431.

[17] Hao Fu, Akshaj Kumar Veldanda, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. 2022. A feature-based on-line detector to remove adversarial-backdoors by iterative demarcation. *IEEE Access* (2022), 5545–5558.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

[19] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[22] Sihong He, Shuo Han, and Fei Miao. 2023. Robust electric vehicle balancing of autonomous mobility-on-demand system: A multi-agent reinforcement learning approach. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5471–5478.

[23] Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. 2023. Robust multi-agent reinforcement learning with state uncertainty. *arXiv preprint arXiv:2307.16212* (2023).

[24] Sihong He, Yue Wang, Shuo Han, Shaofeng Zou, and Fei Miao. 2023. A Robust and Constrained Multi-Agent Reinforcement Learning Electric Vehicle Rebalancing Method in AMoD Systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5637–5644.

[25] Yi He, Fudong Lin, Xu Yuan, and Nian-Feng Tzeng. 2021. Interpretable Minority Synthesis for Imbalanced Classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. 2542–2548.

[26] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural Adversarial Examples. In *CVPR*.

[27] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. 2021. Rethinking Spatial Dimensions of Vision Transformers. In *International Conference on Computer Vision (ICCV)*.

[28] Tiechuan Hu, Wenbo Zhu, and Yuqi Yan. 2023. Artificial intelligence aspect of transportation analysis using large scale systems. In *Artificial Intelligence and Cloud Computing Conference*. 54–59.

[29] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. 2022. LAS-AT: Adversarial Training with Learnable Attack Strategy. In *CVPR*.

[30] Hoki Kim, Woojin Lee, and Jaewook Lee. 2021. Understanding Catastrophic Overfitting in Single-step Adversarial Training. In *AAAI*.

[31] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems (NeurIPS)*.

[33] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. In *ICLR*.

[34] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*.

[35] Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. 2022. Subspace Adversarial Training. In *CVPR*.

[36] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Computer Vision and Pattern Recognition (CVPR)*.

[37] Yichen Li and Chicheng Zhang. 2023. Ensemble-based Interactive Imitation Learning. *arXiv preprint arXiv:2312.16860* (2023).

[38] Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, et al. 2023. MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 5774–5784.

[39] Fudong Lin, Kaleb Guillot, Summer Crawford, Yihe Zhang, Xu Yuan, and Nian-Feng Tzeng. 2024. An Open and Large-Scale Dataset for Multi-Modal Climate Change-aware Crop Yield Predictions. *arXiv preprint arXiv:2406.06081* (2024).

[40] Fudong Lin, Xu Yuan, Lu Peng, and Nian-Feng Tzeng. 2022. Cascade Variational Auto-Encoder for Hierarchical Disentanglement. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*. 1248–1257.

[41] Fudong Lin, Xu Yuan, Yihe Zhang, Purushottam Sigdel, Li Chen, Lu Peng, and Nian-Feng Tzeng. 2023. Comprehensive Transformer-Based Model Architecture for Real-World Storm Prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. 54–71.

[42] Han Liu, Yuhao Wu, Zhiyuan Yu, Yevgeniy Vorobeychik, and Ning Zhang. 2023. Slowlidar: Increasing the latency of lidar-based detection using adversarial examples. In *CVPR*. 5146–5155.

[43] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. 2023. RIATIG: Reliable and Imperceptible Adversarial Text-to-Image Generation with Natural Prompts. In *CVPR*. 20585–20594.

[44] Han Liu, Zhiyuan Yu, Mingming Zha, XiaoFeng Wang, William Yeoh, Yevgeniy Vorobeychik, and Ning Zhang. 2022. When evil calls: Targeted adversarial voice over ip network. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2009–2023.

[45] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. 2022. Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack. In *CVPR*.

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *ICCV*.

[47] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *ICLR*.

[48] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

[49] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fushen Wang, and Chao Chen. 2022. A Multimodal Transformer: Fusing Clinical Notes with Structured EHR Data for Interpretable In-Hospital Mortality Prediction. In *American Medical Informatics Association Annual Symposium (AMIA)*.

[50] Weimin Lyu, Songzhu Zheng, Tengfei Ma, and Chao Chen. 2022. A Study of the Attention Abnormality in Trojaned BERTs. In *NAACL*. 4727–4741.

[51] Shiqing Ma and Yingqi Liu. 2019. Nic: Detecting adversarial samples with neural network invariant checking. In *Network and Distributed System Security Symposium (NDSS)*.

[52] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

[53] Yifei Min, Lin Chen, and Amin Karbasi. 2021. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence*.

[54] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*.

[55] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. 2016. The Limitations of Deep Learning in Adversarial Settings. In *European Symposium on Security and Privacy (EuroS&P)*.

[56] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. 2021. Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints. In *NeurIPS*.

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* (2020).

[58] Jayaram Raghuram, Varun Chandrasekaran, Somesh Jha, and Suman Banerjee. 2021. A General Framework For Detecting Anomalous Inputs to DNN Classifiers. In *ICML*.

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

[60] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. 2019. The Odds are Odd: A Statistical Test for Detecting Adversarial Examples. In *ICML*.

[61] Kangrui Ruan and Xuan Di. 2022. Learning human driving behaviors with sequential causal imitation learning. In *AAAI*. 4583–4592.

[62] Kangrui Ruan and Xuan Di. 2024. InfoSTGCAN: An Information-Maximizing Spatial-Temporal Graph Convolutional Attention Network for Heterogeneous Human Trajectory Prediction. *Computers* 13, 6 (2024), 151.

[63] Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. 2023. Causal imitation learning via inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*.

[64] Ruslan Salakhutdinov and Geoffrey E. Hinton. 2007. Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure. In *AISTATS*.

[65] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*.

[66] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free!. In *NeurIPS*.

[67] Fatemeh Sheikholeslami, Ali Lotfi, and J. Zico Kolter. 2021. Provably robust classification of adversarial examples with detection. In *ICLR*.

[68] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, Yoshua Bengio and Yann LeCun (Eds.).

[69] Han Song, Cong Liu, and Huafeng Dai. 2024. BundledSLAM: An Accurate Visual SLAM System Using Multiple Cameras. In *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Vol. 7. 106–111.

[70] Han Song, Zhongche Qu, Zhi Zhang, Zihan Ye, and Cong Liu. 2024. ETA-INIT: Enhancing the Translation Accuracy for Stereo Visual-Inertial SLAM Initialization. *arXiv preprint arXiv:2405.15082* (2024).

[71] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshop*.

[72] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *NeurIPS*.

[73] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *ICML*.

[74] Florian Tramèr. 2022. Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them. In *ICML*.

[75] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On Adaptive Attacks to Adversarial Example Defenses. In *NeurIPS*.

[76] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.

[77] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research (JMLR)* (2008).

[78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*.

[79] Haotao Wang, Aston Zhang, Shuai Zheng, Xingjian Shi, Mu Li, and Zhangyang Wang. 2022. Removing Batch Normalization Boosts Adversarial Training. In *ICML*.

[80] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *ICCV*.

[81] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. PVT v2: Improved baselines with Pyramid Vision Transformer. *Comput. Vis. Media* (2022).

[82] Eric Wong, Leslie Rice, and J. Zico Kolter. 2020. Fast is better than free: Revisiting adversarial training. In *ICLR*.

[83] Jun Wu, Xuesong Ye, and Yanyuet Man. 2023. Bottrinet: A unified and efficient embedding for social bots detection via metric learning. In *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*. 1–6.

[84] Jun Wu, Xuesong Ye, and Chengjie Mou. 2023. Botshape: A Novel Social Bots Detection Approach Via Behavioral Patterns. In *International Conference on Data Mining & Knowledge Management Process*.

[85] Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2020. ML-LOO: Detecting Adversarial Examples with Feature Attribution. In *AAAI*.

[86] Chengyuan Yao, Pavol Bielik, Petar Tsankov, and Martin T. Vechev. 2021. Automated Discovery of Adaptive Attacks on Adversarial Defenses. In *NeurIPS*.

[87] Chengyuan Yao, Pavol Bielik, Petar Tsankov, and Martin T. Vechev. 2021. Automated Discovery of Adaptive Attacks on Adversarial Defenses. In *NeurIPS*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.).

[88] Xuwang Yin, Soheil Kolouri, and Gustavo K. Rohde. 2020. GAT: Generative Adversarial Training for Adversarial Example Detection and Robust Classification. In *ICLR*.

[89] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. In *International Conference on Computer Vision (ICCV)*.

[90] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*.

[91] Mingkang Zhu, Tianlong Chen, and Zhangyang Wang. 2021. Sparse and Imperceptible Adversarial Attack via a Homotopy Algorithm. In *ICML*.

[92] Wenbo Zhu and Tiechuan Hu. 2021. Twitter Sentiment analysis of covid vaccines. In *International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. 118–122.

[93] Jun Zhuang and Mohammad Al Hasan. 2022. Defending graph convolutional networks against dynamic graph perturbations via bayesian self-supervision. In *AAAI*. 4405–4413.

[94] Jun Zhuang and Mohammad Al Hasan. 2022. Robust node classification on graphs: Jointly from bayesian label transition and topology-based label propagation. In *International Conference on Information & Knowledge Management*. 2795–2805.