Be Your Own Neighborhood: Detecting Adversarial Examples by the Neighborhood Relations Built on Self-Supervised Learning

Zhiyuan He [*1] **Yijun Yang** [*1] **Pin-Yu Chen** [2] **Qiang Xu** [1] **Tsung-Yi Ho** [1]

Abstract

Deep Neural Networks (DNNs) are vulnerable to Adversarial Examples (AEs), hindering their use in safety-critical systems. In this paper, we present **BEYOND**, an innovative AE detection framework designed for reliable predictions. BEYOND identifies AEs by distinguishing the AE's abnormal relation with its augmented versions, i.e. neighbors, from two prospects: representation similarity and label consistency. An off-the-shelf Self-Supervised Learning (SSL) model is used to extract the representation and predict the label for its highly informative representation capacity compared to supervised learning models. We found clean samples maintain a high degree of representation similarity and label consistency relative to their neighbors, in contrast to AEs which exhibit significant discrepancies. We explain this observation and show that leveraging this discrepancy BEYOND can accurately detect AEs. Additionally, we develop a rigorous justification for the effectiveness of BEYOND. Furthermore, as a plug-and-play model, BEYOND can easily cooperate with the Adversarial Trained Classifier (ATC), achieving state-of-the-art (SOTA) robustness accuracy. Experimental results show that BEYOND outperforms baselines by a large margin, especially under adaptive attacks. Empowered by the robust relationship built on SSL, we found that BEYOND outperforms baselines in terms of both detection ability and speed. Project page: https://huggingface.co/spaces/allenhzy/Be-Your-Own-Neighborhood.

[*]Equal contribution [1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Sha Tin, Hong Kong. yjyang@cse.cuhk.edu.hk, qxu@cse.cuhk.edu.hk, tyho@cse.cuhk.edu.hk. [2]IBM Research, New York, USA. pin-yu.chen@ibm.com. Correspondence to: Zhiyuan He <zyhe@cse.cuhk.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Deep Neural Networks (DNNs) have been widely adopted in many fields due to their superior performance. However, their susceptibility to Adversarial Examples (AEs), which can easily fool DNNs by adding some imperceptible adversarial perturbations, limits their deployment in safety-critical scenarios such as autonomous driving (Cococcioni et al., 2020) and disease diagnosis (Kaissis et al., 2020), where incorrect predictions can lead to catastrophic economic and even loss of life.

Existing defensive strategies can be roughly categorized as adversarial training, input purification (Mao et al., 2021), and AE detection (Xu et al., 2017). Adversarial training is known as the most effective defense technique (Croce & Hein, 2020), but it brings degradation of accuracy and additional training costs, which are unacceptable in some application scenarios. In contrast, input purification techniques avoid these costs, but their defensive ability is limited, i.e. easily defeated by adaptive attacks (Croce & Hein, 2020).

Recently, a large number of AE detection methods have been proposed (Zuo & Zeng, 2021). Some methods detect AE by interrogating the abnormal relationship between AE and other samples. For example, Deep k-Nearest Neighbors (DkNN) (Papernot & McDaniel, 2018) compares the DNN-extracted features of the input image with those of its k nearest neighbors layer by layer to identify AEs, leading to a high inference time. Latent Neighborhood Graph (LNG) (Abusnaina et al., 2021) represents the relationship between the input sample and the reference sample as a graph, whose nodes are embeddings extracted by DNN and edges are built according to distances between the input node and reference nodes, and train a graph neural network to detect AEs.

Though more efficient than DkNN, LNG suffers from some weaknesses: some AEs are required to build the graph, so its detection performance relies on the reference AEs and cannot effectively generalize to unseen attacks. More importantly, both DkNN and LNG can be bypassed by adaptive attacks, in which the adversary has full knowledge of the detection strategy.

We observe that one cause for adversarial vulnerability is the

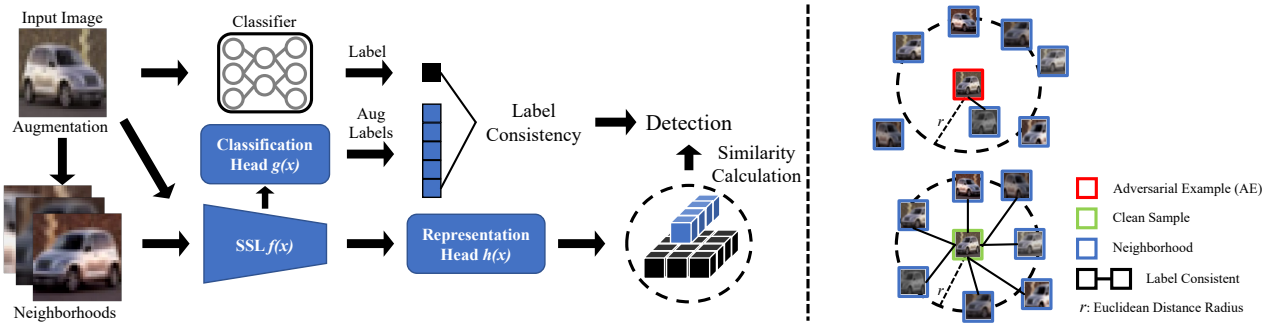


Figure 1. Pipeline of the proposed **BEYOND** framework. First, we augment the input image to obtain a bunch of its neighbors. Then, we perform the label consistency detection mechanism on the classifier's prediction of the input image and that of neighbors predicted by SSL's classification head. Meanwhile, the representation similarity mechanism employs *cosine similarity* to measure the similarity among the input image and its neighbors (left). The input image with poor label consistency or representation similarity is flagged as AE (right).

lack of feature invariance (Jiang et al., 2020), i.e., small perturbations may lead to undesired large changes in features or even predicted labels. On the other hand, Self-Supervised Learning (SSL) (Chen & He, 2021) models learn data representation consistency under different data augmentations, which intuitively can mitigate the issue of lacking feature invariance and thus improve adversarial robustness.

To clarify our findings, we visualize the SSL-extracted representation of the clean sample, AE, and that of their corresponding augmentations in Fig. 1 (right). It is evident that clean samples exhibit a stronger correlation with their neighbors in terms of label consistency and representation similarity. In contrast, AEs are distinctly separated from their neighbors.

Inspired by the above observations, we propose a novel AE detection framework, named **BE Your Own NeighborhooD (BEYOND)**. The contributions of this work are summarized as follows:

- We propose BEYOND, a novel AE detection framework, which utilizes the robust representation capacity of SSL model to identify AE by examining their proximity to neighbor samples generated by augmentations. To our best knowledge, BEYOND is the first work that leverages an SSL model for AE detection without prior knowledge of adversarial attacks or AEs.
- We develop a rigorous justification for the effectiveness of BEYOND against adversarial and adaptive attacks.
- BEYOND can defend effectively against adaptive attacks. To defeat the two detection mechanisms: label consistency and representation similarity simultaneously, attackers have to optimize two objectives with contradictory directions, resulting in gradients canceling each other out.
- As a plug-and-play method, BEYOND can be applied directly to any image classifier without compromising accuracy or additional retraining costs.

Experimental results show that BEYOND outperforms baselines by a large margin, especially under adaptive attacks. Empowered by the robust relation net built on SSL, we found BEYOND outperforms baselines in terms of both detection ability and implementation costs.

2. Related Works

The authors in (Szegedy et al., 2013) first discovered that an adversary could maximize the prediction error of the network by adding some imperceptible perturbation, δ, which is typically bounded by a perturbation budget, ϵ, under an L_p-norm, e.g., L_∞ and L_2. Project Gradient Descent (PGD) proposed by (Madry et al., 2017) is one of the most powerful iterative attacks. PGD motivates various gradient-based attacks such as AutoAttack (Croce & Hein, 2020) and Orthogonal-PGD (Bryniarski et al., 2021), which can break many SOTA AE defenses (Croce et al., 2022). Another widely adopted adversarial attack is C&W (Carlini & Wagner, 2017). Compared to the norm-bounded PGD attack, C&W conducts AEs with a high attack success rate by formulating the adversarial attack problem as an optimization problem.

Existing defense techniques focus either on robust prediction or detection. The most effective way to achieve robust prediction is adversarial training (Elfwing et al., 2018; Zhang et al., 2019), and the use of nearest neighbors is a common approach to detecting AEs. kNN (Dubey et al., 2019) and DkNN (Papernot & McDaniel, 2018) discriminate AEs by checking the label consistency of each layer's neighborhoods. (Ma et al., 2018) define Local Intrinsic Dimensionality (LID) to characterize the properties of AEs and use a simple k-NN classifier to detect AEs. LNG (Abusnaina

et al., 2021) searches for the nearest samples in the reference data and constructs a graph, further training a specialized GNN to detect AEs. Although these nearest-neighbor-based methods achieve competitive detection performance, all rely on external AEs for training detectors or searching thresholds, resulting in defeat against unseen attacks.

Recent studies have shown that SSL can improve adversarial robustness as SSL models are label-independent and insensitive to transformations (Hendrycks et al., 2019). An intuitive idea is to combine adversarial training and SSL (Ho & Nvasconcelos, 2020; Kim et al., 2020), which remain computationally expensive and not robust to adaptive attacks. (Shi et al., 2021) and (Mao et al., 2021) find that the auxiliary SSL task can be used to purify AEs, which are shown to be robust to adaptive attacks. However, (Croce et al., 2022) shows these adaptive test-time defenses can be broken by stronger adaptive attacks.

3. BEYOND: Proposed Method

This section provides a detailed explanation of the proposed BEYOND. We begin by outlining the core components of the BEYOND design and elaborating on the detection algorithm. Following this, we present a theoretical justification for BEYOND's effectiveness against both grey-box and adaptive attacks.

3.1. Method Overview

Components. BEYOND consists of three components: a SSL feature extractor $f(\cdot)$, a classification head $g(\cdot)$, and a representation head $h(\cdot)$, as shown in Fig. 1 (left). Specifically, the SSL feature extractor is a Convolutional Neural Network (CNN), pre-trained by specially designed loss, e.g. contrastive loss, without supervision[1]. A Fully-Connected layer (FC) acts as the classification head $g(\cdot)$, trained by freezing the $f(\cdot)$. The $g(\cdot)$ performs on the input image's neighbors for label consistency detection. The representation head $h(\cdot)$ consisting of three FCs, encodes the output of $f(\cdot)$ to an embedding, i.e. representation. We operate the representation similarity detection between the input image and its neighbors.

Core idea. Our approach relies on robust relationships between the input and its neighbors for the detection of AE. The key idea is that adversaries may easily attack one sample's representation to another submanifold, but it is difficult to totally shift that of all its neighbors. We employ the SSL model to capture such relationships since it is trained to project input and its augmentations (neighbors) to the same submanifold (Chen & He, 2021).

[1]Here, we employ the SimSiam (Chen & He, 2021) as the SSL feature-extractor for its decent performance.

Algorithm 1 BEYOND detection algorithm

Input: Input image x, target classifier $c(\cdot)$, SSL feature extractor $f(x)$, classification head $g(x)$, projector head $h(x)$, label consistency threshold $\mathcal{T}_{label}$, cosine similarity threshold $\mathcal{T}_{cos}$, representation similarity threshold $\mathcal{T}_{rep}$, Augmentation Aug, neighbor indicator i, total neighbor k
Output: `reject / accept`

Stage1: Collect labels and representations.
$\ell_{cls}(x) = c(x)$
for i in k **do**
  $\hat{x}_i = Aug(x)$
  $\ell_{ssl}(\hat{x}_i) = f(g(\hat{x}_i))$;$r(x) = f(h(x))$; $r(\hat{x}_i) = f(h(\hat{x}_i))$
Stage2: Label consistency detection mechanism.
for i in k **do**
  if $\ell(\hat{x}_i) == \ell(x)$ **then** $\texttt{Ind}_{\texttt{label}} += 1$
Stage3: Representation similarity detection mechanism.
for i in k **do**
if $cos(r(x), r(\hat{x}_i)) < \mathcal{T}_{cos}$ **then** $\texttt{Ind}_{rep} += 1$
Stage4: AE detection.
if $\texttt{Ind}_{label} < \mathcal{T}_{label}$ or $\texttt{Ind}_{rep} < \mathcal{T}_{rep}$ **then** `reject`
else `accept`

Selection of neighbor number. Obviously, the larger the number of neighbors, the more stable the relationship between them, but this may increase the overhead. We choose 50 neighbors for BEYOND, since larger neighbors no longer significantly enhance performance, as shown in Fig. 3.

Workflow. Fig. 1 shows the workflow of the proposed BEYOND. When input comes, we first transform it into 50 augmentations, i.e. 50 neighbors. Note that BEYOND is not based on random data augmentation. Next, the input along with its 50 neighbors are fed to SSL feature extractor $f(\cdot)$ and then the classification head $g(\cdot)$ and the representation head $h(\cdot)$, respectively. For the classification branch, $g(\cdot)$ outputs the predicted label for 50 neighbors. Later, the label consistency detection algorithm calculates the consistency level between the input label (predicted by the classifier) and 50 neighbor labels. When it comes to the representation branch, the 51 generated representations are sent to the representation similarity detection algorithm for AE detection. If the consistency of the label of a sample or its representation similarity is lower than a threshold, BEYOND shall flag it AE.

3.2. Detection Algorithms

For enhanced AE detection capability, BEYOND adopted two detection mechanisms: *Label Consistency*, and *Representation Similarity*. The detection performance of the two combined can exceed any of the individuals. More importantly, their contradictory optimization directions hinder adaptive attacks to bypass both of them simultaneously.

Label Consistency. We compare the classifier prediction, $\ell_{cls}(x)$, on the input image, x, with the predictions of the SSL classification head, $\ell_{ssl}(\hat{x}_i), i = 1 \ldots k$, where $\hat{x}_i$ de-

notes the ith neighbor, k is the total number of neighbors. If $\ell_{cls}(x)$ equals $\ell_{ssl_i}(\hat{x}_i)$, the label consistency increases by one, $\texttt{Ind}_{\texttt{Label}} += 1$. Once the final label consistency is less than the threshold, $\texttt{Ind}_{\texttt{Label}} < \mathcal{T}_{label}$, the *Label Consistency* flags it as AE. We summarize the label consistency detection mechanism in Algorithm. 1.

Representation Similarity. We employ the *cosine distance* as a metric to calculate the similarity between the representation of input sample $r(x)$ and that of its neighbors, $r(\hat{x}_i), i = 1, ..., k$. Once the similarity, $cos(r(x), r(\hat{x}_i))$, is smaller than a certain value, representation similarity increases by 1, $\texttt{Ind}_{\texttt{Rep}} += 1$. If the final representation similarity is less than a threshold, $\texttt{Ind}_{\texttt{ReP}} < \mathcal{T}_{rep}$, the *representation similarity* flag the sample as an AE. Algorithm. 1 concludes the representation similarity detection mechanism.

Note that, we select the thresholds, i.e. $\mathcal{T}_{label}$, $\mathcal{T}_{rep}$, by fixing the False Positive Rate (FPR)@5%, which can be determined only by clean samples, and the implementation of our method needs no prior knowledge about AE.

4. Theoretical Justification

4.1. Theoretical Analysis

Given a clean sample x, we receive its feature $f(x)$ lying in the feature space spanned by the SSL model. We assume that benign perturbation, i.e. random noise, $\hat{\delta}$, with bounded budgets causes minor variation, $\hat{\varepsilon}$, on the feature space, as described in Eq. 1:

$$f(x + \hat{\delta}) = f(x) + \nabla f(x)\hat{\delta} = f(x) + \hat{\varepsilon}, \tag{1}$$

where $\|\hat{\varepsilon}\|_2$ is constrained to be within a radius r. In contrast, when it comes to AE, x_{adv}, the adversarial perturbation, δ, can cause considerable change, due to its maliciousness, that is, it causes misclassification and transferability (Demontis et al., 2019; Liu et al., 2021; Papernot et al., 2016), as formulated in Eq. 2.

$$f(x_{adv}) = f(x + \delta) = f(x) + \nabla f(x)\delta = f(x) + \varepsilon, \tag{2}$$

where $\|\varepsilon\|_2$ is significantly larger than $\|\hat{\varepsilon}\|_2$ formally, $\lim_{\hat{\varepsilon}\to 0} \frac{\varepsilon}{\hat{\varepsilon}} = \infty$. SSL model is trained to generate close representations between an input x and its augmentation $x_{aug} = Wx$ (Hendrycks et al., 2019; Jaiswal et al., 2020), where $W \in \mathbb{R}^{w\times h}$, w, h denote the width and height of x, respectively. Based on this natural property of SSL ($f(Wx) \approx f(x)$), we have:

$$f(Wx) = f(x) + o(\hat{\varepsilon}), \nabla f(Wx) = \nabla f(x) + o(\hat{\varepsilon}), \tag{3}$$

where $o(\hat{\varepsilon})$ is a high-order infinitesimal item of $\hat{\varepsilon}$. Moreover, according to Eq. 1 and Eq. 3, we can derive that:

$$\begin{aligned} f(W(x + \hat{\delta})) &= f(Wx) + \nabla f(Wx)W\hat{\delta} \\ &= f(x) + \nabla f(x)W\hat{\delta} + o(\hat{\varepsilon}). \end{aligned} \tag{4}$$

We let $\hat{\varepsilon}_{aug} = \nabla f(x)W\hat{\delta}$ and assume $\hat{\varepsilon}_{aug}$ and $\hat{\varepsilon}$ are infinitesimal isotropic, i.e. $\lim_{\hat{\varepsilon}\to 0} \frac{\hat{\varepsilon}_{aug}}{\hat{\varepsilon}} = c$, where c is a constant. Therefore, we can rewrite Eq. 4 as follows:

$$f(W(x + \hat{\delta})) = f(x) + c \cdot \hat{\varepsilon} + o(\hat{\varepsilon}). \tag{5}$$

Our goal is to prove that *distance (similarity) between AE and its neighbors can be significantly larger (smaller) than that of the clean sample in the space spanned by a SSL model*, which is equivalent to justify Eq. 6:

$$\|f(x_{adv}) - f(W(x_{adv}))\|_2^2 \geq \|\underbrace{f(x) - f(Wx)}_{\hat{\varepsilon}_{aug}=c\cdot\hat{\varepsilon}}\|_2^2. \tag{6}$$

Expending the left-hand item in Eq. 6, and defining $m = \nabla f(x)W\delta$, we can obtain the following.

$$\begin{aligned} &\|f(x_{adv}) - f(Wx_{adv})\|_2^2 = \|f(x + \delta) - f(W(x + \delta))\|_2^2 \\ &= \|f(x) + \nabla f(x)\delta - f(Wx) - \nabla f(Wx)W\delta\|_2^2 \\ &= \|\varepsilon - \nabla f(x)W\delta - o(\varepsilon)\|_2^2 \\ &= \|\varepsilon\|_2^2 + \|m\|_2^2 - 2|\langle\varepsilon, m\rangle| + o(\varepsilon) \end{aligned} \tag{7}$$

As mentioned in the prior literature (Mikołajczyk & Grochowski, 2018; Raff et al., 2019; Zeng et al., 2020), augmentations can effectively weaken adversarial perturbation δ. Therefore, we assume that the influence caused by $W\delta$ is weaker than δ but stronger than the benign perturbation, $\hat{\delta}$. Formally, we have:

$$\|\underbrace{\nabla f(x)\delta}_{\varepsilon}\|_2 > \|\underbrace{\nabla f(x)W\delta}_{m}\|_2 > \|\underbrace{\nabla f(x)W\hat{\delta}}_{\hat{\varepsilon}_{aug}=c\cdot\hat{\varepsilon}}\|_2. \tag{8}$$

According to *Cauchy–Schwarz inequality* (Bhatia & Davis, 1995), we have the following chain of inequalities obtained by taking Eq. 8 into Eq. 7:

$$\begin{aligned} &\|\varepsilon\|_2^2 + \|m\|_2^2 - 2|\langle\varepsilon, m\rangle| + o(\varepsilon) > \\ &\|\varepsilon\|_2^2 + \|m\|_2^2 - 2\|\varepsilon\| \cdot \|m\| = (\|\varepsilon\|_2 - \|m\|_2)^2, \end{aligned} \tag{9}$$

where $\|m\| \in (\|\hat{\varepsilon}_{aug}\|, \|\varepsilon\|)$ according to Eq. 8.

Finally, from Eq. 9 we observe that by applying proper data augmentation, the distance between AE and its neighbors in SSL's feature space $\|f(x_{adv}) - f(Wx_{adv})\|_2 = \|\|\varepsilon\|_2 - \|m\|_2\|_2$ can be significantly larger than that of clean samples $\|f(x) - f(Wx)\|_2 = o(\hat{\varepsilon})$. The enlarged distance is upper bounded by $\|\varepsilon\|_2/\|\hat{\varepsilon}_{aug}\|_2$ times that of the clean sample, which implies that the imperceptible perturbation δ in the image space can be significantly enlarged in SSL's feature space by referring to its neighbors. This exactly supports the design of BEYOND as described in Sec 3.1. In practice, we adopt various augmentations instead of a single type to generate multiple neighbors for AE detection, which reduces the randomness, resulting in more robust estimations.

Table 1. The AUC of Different Adversarial Detection Approaches on CIFAR-10. The results are the mean and standard deviation of 5 runs. LNG is not open-sourced and the data comes from its report. To align with baselines, classifier: ResNet110, FGSM: $\epsilon = 0.05$, PGD: $\epsilon = 0.02$. Note that **BEYOND needs no AE for training**, leading to the same value on both *seen* and *unseen* settings. The **bolded** values are the best performance, and the *underlined italicized* values are the second-best performance, the same below.

AUC (%)	*Unseen*: Attacks used in training are preclude from tests.				*Seen*: Attacks used in training are included in tests.				
	FGSM	PGD	AutoAttack	Square	FGSM	PGD	CW	AutoAttack	Square
DkNN	$61.55_{\pm 0.023}$	$51.22_{\pm 0.026}$	$52.12_{\pm 0.023}$	$59.46_{\pm 0.022}$	$61.55_{\pm 0.023}$	$51.22_{\pm 0.026}$	$61.52_{\pm 0.028}$	$52.12_{\pm 0.023}$	$59.46_{\pm 0.022}$
kNN	$61.83_{\pm 0.018}$	$54.52_{\pm 0.022}$	$52.67_{\pm 0.022}$	$73.39_{\pm 0.02}$	$61.83_{\pm 0.018}$	$54.52_{\pm 0.022}$	$62.23_{\pm 0.019}$	$52.67_{\pm 0.022}$	$73.39_{\pm 0.02}$
LID	$71.08_{\pm 0.024}$	$61.33_{\pm 0.025}$	$55.56_{\pm 0.021}$	$66.18_{\pm 0.025}$	$73.61_{\pm 0.02}$	$67.98_{\pm 0.02}$	$55.68_{\pm 0.021}$	$56.33_{\pm 0.024}$	$85.94_{\pm 0.018}$
Hu	$84.51_{\pm 0.025}$	$58.59_{\pm 0.028}$	$53.55_{\pm 0.029}$	*$95.82_{\pm 0.02}$*	$84.51_{\pm 0.025}$	$58.59_{\pm 0.028}$	*$91.02_{\pm 0.022}$*	$53.55_{\pm 0.029}$	$95.82_{\pm 0.02}$
Mao	$95.33_{\pm 0.012}$	*$82.61_{\pm 0.016}$*	*$81.95_{\pm 0.02}$*	$85.76_{\pm 0.019}$	$95.33_{\pm 0.012}$	$82.61_{\pm 0.016}$	$83.10_{\pm 0.018}$	$81.95_{\pm 0.02}$	$85.76_{\pm 0.019}$
LNG	*98.51*	63.14	58.47	94.71	**99.88**	*91.39*	89.74	*84.03*	*98.82*
BEYOND	**98.89**$_{\pm 0.013}$	**99.28**$_{\pm 0.02}$	**99.16**$_{\pm 0.021}$	**99.27**$_{\pm 0.016}$	*98.89*$_{\pm 0.013}$	**99.28**$_{\pm 0.02}$	**99.20**$_{\pm 0.008}$	**99.16**$_{\pm 0.021}$	**99.27**$_{\pm 0.016}$

4.2. Robustness to Adaptive Attacks

Adaptive Objective Loss Function. Attackers can design adaptive attacks to try to bypass BEYOND when the attacker knows all the parameters of the model and the detection strategy. For an SSL model with a feature extractor f, a projector h, and a classification head g, the classification branch can be formulated as $\mathbb{C} = f \circ g$ and the representation branch as $\mathbb{R} = f \circ h$. To attack effectively, the adversary must deceive the target model while guaranteeing the label consistency and representation similarity of the SSL model. Since BEYOND uses multiple augmentations, we estimate their impact on label consistency and representation similarity during the adaptive attack following Expectation over Transformation (EoT) (Athalye et al., 2018b) as:

$$\begin{aligned} Sim_l &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}\left(\mathbb{C}\left(W^i(x+\delta)\right), y_t\right) \\ Sim_r &= \frac{1}{k}\sum_{i=1}^{k} \mathcal{S}(\mathbb{R}(W^i(x+\delta)), \mathbb{R}(x+\delta)) \end{aligned} \tag{10}$$

where $\mathcal{S}$ represents cosine similarity, k represents the number of generated neighbors, and the linear augmentation function $W(x) = W(x, p)$; $p \sim P$ randomly samples p from the parameter distribution P to generate different neighbors. Note that we guarantee the generated neighbors are fixed each time by fixing the random seed. The adaptive adversaries perform attacks on the following objective function:

$$\min_{\delta} \mathcal{L}_C(x+\delta, y_t) + Sim_l - \alpha \cdot Sim_r, \tag{11}$$

where $\mathcal{L}_C$ indicates classifier's loss function, y_t is the targeted class, and α refers to a hyperparameter[2], which is a trade-off parameter between label consistency and representation similarity. Experiments in the Appendix show that the adaptive attack is most effective when $\alpha = 1$.

Conflicting Optimization Goals. For an AE $x_{adv} = x + \delta$ and $y_{adv} = \mathbb{C}(x_{adv})$, the classification and representation outputs of its augmentation can be studied through their first-order Taylor expansion at x:

$$\begin{aligned} y_{aug} &= \mathbb{C}(Wx_{adv}) = \mathbb{C}(Wx) + \nabla\mathbb{C}(Wx)W\delta \\ r_{aug} &= \mathbb{R}(Wx_{adv}) = \mathbb{R}(Wx) + \nabla\mathbb{R}(Wx)W\delta \end{aligned} \tag{12}$$

Since the SSL model is trained to generate close representations between a sample and its augmentation ($\mathbb{C}(Wx) \approx \mathbb{C}(x), \mathbb{R}(Wx) \approx \mathbb{R}(x)$), the differences of label and representation between the original sample and its augmentation are denoted as:

$$\begin{aligned} y_{aug} - y &\approx \nabla\mathbb{C}(x)W\delta \\ r_{aug} - r &\approx \nabla\mathbb{R}(x)W\delta \end{aligned} \tag{13}$$

Therefore, to ensure the label consistency of AE, i.e., $y_{aug} = y_t \neq y$, the optimization goal of the adaptive attack is making δ larger within the perturbation budget:

$$\delta = \underset{\|\delta\| \leq \epsilon}{\arg\max} \left(\nabla\mathbb{C}(x)W\delta\right) \tag{14}$$

Conversely, the optimization goal of representation similarity ($r_{aug} = r$) is making δ smaller:

$$\delta = \underset{\|\delta\| \leq \epsilon}{\arg\min} \left(\nabla\mathbb{R}(x)W\delta\right) \tag{15}$$

Since the classification $\mathbb{C}$ and representation head $\mathbb{R}$ share the same backbone f, optimizing for these conflicting goals can lead to gradient cancellation, which underpins the robustness of BEYOND against adaptive attacks. Fig. 10 visualizes the gradient sign associated with these objectives, showing the phenomenon of gradient cancellation due to conflicting goals.

Moreover, the above analysis demonstrates that small perturbations do not guarantee label consistency for AEs, while large perturbations impair representation similarity, which is consistent with the empirical results in Sec 5.4.

5. Evaluation

This section details the experimental setting used to evaluate the performance of BEYOND. We outline the datasets, target

[2]Note that we employ cosine metric that is negatively correlated with the similarity, so that the Sim_r item is preceded by a minus sign.

models, attack methods, evaluation metrics, and baseline methods employed in our experiments. Furthermore, we present the results of BEYOND's performance against both limited knowledge and perfect knowledge attacks.

5.1. Experimental Setting

Limited knowledge attack & Perfect knowledge attack. Following (Apruzzese et al., 2023), in the limited knowledge attack setting, the adversary has complete knowledge of the classifier, while the detection strategy is confidential. Whereas in an adaptive attack (perfect knowledge) setting, the adversary is aware of the detection strategy.

Datasets & Target models. We conduct experiments on three commonly adopted datasets including CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100, and IMAGENET (Krizhevsky et al., 2012) The details of the target models (classifiers), and the employed SSL models together with their original classification accuracy on clean samples are summarized in Table 8 [3].

Augmentations. The types of augmentation used by BEYOND to generate neighbors are consistent with Sim-Siam, including horizontal flipping, cropping, color jitter, and greyscale. However, BEYOND fixes the random seed to prevent benefiting from randomization. We generate 50 neighbors for each sample, and the ablation study on the number of neighbors is further discussed in Sec. 5.4.

Attacks. Evaluations of limited knowledge attacks are conducted on FGSM, PGD, C&W, and AutoAttack. AutoAttack includes APGD, APGD-T, FAB-T, and Square (Andriushchenko et al., 2020), where APGD-T and FAB-T are targeted attacks and Square is a black-box attack. As for adaptive attacks, we employed the most adopted EoT and Orthogonal-PGD, which is a recent adaptive attack designed for AE detectors.

Metrics. Following previous work (Abusnaina et al., 2021), we employ ROC curve & AUC and Robust Accuracy (RA) as evaluation metrics.

- **ROC curve & AUC:** Receiver Operating Characteristic (ROC) curves describe the impact of various thresholds on detection performance, and the Area Under the Curve (AUC) is an overall indicator of the ROC curve.
- **Robust Accuracy (RA)**: We employ RA as an evaluation metric, which can reflect the overall system performance against adaptive attacks by considering both the classifier and the detector.

Baselines. We choose five detection-based defense methods as baselines: kNN (Dubey et al., 2019), DkNN (Papernot & McDaniel, 2018), LID (Ma et al., 2018), (Hu et al., 2019) and LNG, which also consider the relationship between the input and its neighbors to some extent. (Mao et al., 2021) trains self-supervised branches to purify the adversarial examples, which is one of the best adaptive robust methods available.

[3]The pre-trained SSL models for CIFAR-10 and CIFAR-100 are from Solo-learn (da Costa et al., 2022), and for IMAGENET are from SimSiam (Chen & He, 2021).

Table 2. The AUC of Different Adversarial Detection Approaches on IMAGENET. To align with baselines, classifier: DenseNet121, FGSM: $\epsilon = 0.05$, PGD: $\epsilon = 0.02$. Due to memory and resource constraints, baseline methods are not evaluated against AutoAttack on IMAGENET.

AUC (%)	Unseen		Seen		
	FGSM	PGD	FGSM	PGD	CW
DkNN	$89.16_{\pm 0.038}$	$78.00_{\pm 0.041}$	$89.16_{\pm 0.038}$	$78.00_{\pm 0.041}$	$68.91_{\pm 0.044}$
kNN	$51.63_{\pm 0.04}$	$51.14_{\pm 0.039}$	$51.63_{\pm 0.04}$	$51.14_{\pm 0.039}$	$50.73_{\pm 0.04}$
LID	$90.32_{\pm 0.046}$	$52.56_{\pm 0.038}$	$\underline{99.24}_{\pm 0.043}$	$\underline{98.09}_{\pm 0.042}$	$58.83_{\pm 0.041}$
Hu	$72.56_{\pm 0.037}$	$86.00_{\pm 0.042}$	$72.56_{\pm 0.037}$	$86.00_{\pm 0.042}$	$80.79_{\pm 0.044}$
LNG	$\underline{96.85}$	$\underline{89.61}$	**99.53**	**98.42**	$\underline{86.05}$
BEYOND	$\mathbf{97.59}_{\pm 0.04}$	$\mathbf{96.26}_{\pm 0.045}$	$97.59_{\pm 0.04}$	$96.26_{\pm 0.045}$	$\mathbf{95.46}_{\pm 0.047}$

5.2. Defending Limited Knowledge Attacks

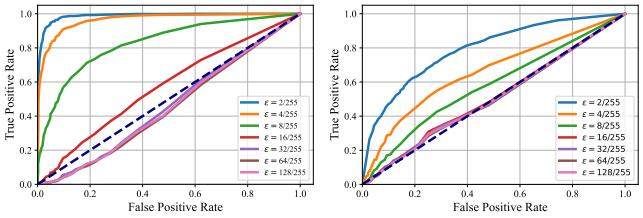
We compare the AUC of BEYOND with DkNN, kNN, LID, Hu, Mao, and LNG on CIFAR-10 and IMAGENET. Since LID and LNG rely on reference AEs, we report detection performance on both seen and unseen attacks. In the seen attack setting, LID and LNG are trained with all types of attacks, while using only the C&W attack in the unseen attack setting. Note that the detection performance for seen and unseen attacks is consistent for detection methods without AEs training.

Table 1 shows that BEYOND consistently surpasses SOTA AE detectors on CIFAR-10, with a pronounced edge in detecting unseen attacks. This superior performance is attributed to BEYOND's innovative use of data augmentations as neighbor samples, which is independent of prior adversarial knowledge. Our analysis in Sec 4 confirms that adversarial perturbations disrupt label consistency and representation similarity, which enables BEYOND to distinguish AEs from benign ones with high accuracy.

Table 2 compares the AUC scores of BEYOND with SOTA AE detectors on IMAGENET. Experimental results show that BEYOND outperforms the SOTA AE detectors in detecting unseen attacks. For seen attacks, the detection performance of BEYOND against FGSM and PGD is marginally lower than that of LNG, which may arise from the fact that prior AEs provide more accurate information on complex datasets. While for stronger attacks, i.e, C&W, BEYOND outperforms baselines by a significant margin. For more information about BEYOND's detection performance (TPR@FPR) on CIFAR-10, CIFAR-100 and IMAGENET, please refer to the Appendix.

Table 3. ATC+BEYOND against AutoAttack on CIFAR-10.

Model	RA		Acc. on clean samples	
	ATC	ATC+BEYOND	ATC	ATC+BEYOND
R2021Fixing70 (Rebuffi et al., 2021)	66.20%	**84.40%**	92.23%	**92.83%**
G2021Improving70 (Gowal et al., 2021)	64.10%	**81.50%**	88.74%	**90.81%**
G2020Uncovering70 (Gowal et al., 2020)	64.70%	**83.80%**	91.10%	**91.79%**
R2021Fixing106 (Rebuffi et al., 2021)	62.20%	**81.30%**	88.50%	**90.51%**



(a) CIFAR-10 (b) IMAGENET

Figure 2. ROC Curve of BEYOND against adaptive attacks with different perturbation budgets.

Improved Robustness with ATC. As a plug-and-play approach, BEYOND integrates well with existing Adversarial Trained Classifier (ATC)[4]. Table 3 shows the accuracy on clean samples and RA against AutoAttack of ATC combined with BEYOND on CIFAR-10. As can be seen the addition of BEYOND increases the robustness of ATC by a significant margin on both clean samples and AEs. Note that the Acc in Table 3 is defined in the Appendix.

5.3. Defending Adaptive Attacks

ROC Curve across Perturbation Budgets. Fig. 2 summarizes the ROC curve varying with different perturbation budgets on CIFAR-10 and IMAGENET. Our analysis regarding Fig. 2 is as follows: 1) BEYOND can be bypassed when perturbations are large enough, due to large perturbations circumventing the transformation. This proves that BEYOND is not gradient masking (Athalye et al., 2018a) and our adaptive attack design is effective. However, large perturbations are easier to perceive. 2) When the perturbation is small, the detection performance of BEYOND for adaptive attacks still maintains a high level, because small perturbations cannot guarantee both label consistency and representation similarity (as shown in Fig. 5 (a)). The above empirical conclusions are consistent with the analysis in Sec 4.2.

Performance against Orthogonal-PGD Adaptive Attacks. Orthogonal-Projected Gradient Descent (Orthogonal-PGD) (Bryniarski et al., 2021) is a cutting-edge benchmark for evaluating the resilience of AE detection methods against adaptive attacks. Orthogonal-PGD has two attack strategies: orthogonal and selection. Table 4 shows BEYOND outperforms the four baselines by a considerable margin in orthogonal strategy, especially under small perturbations. For the worst case, BEYOND can still keep 13.8% ($L_\infty = 8/255$). Furthermore, incorporating ATC can significantly improve the detection performance of BEYOND against large perturbation to 94.5%. See the Appendix for more selection strategy results. In addition, the coupling of the classifier and defense model in Mao's method is not consistent with the Orthogonal-PGD setting. We compare the robust accuracy of BEYOND and Mao for general adaptive attacks in Table 5, which shows that BEYOND outperforms Mao et al. against adaptive attacks with both standard classifier and ATC.

[4]All ATCs are sourced from RobustBench (Croce et al., 2020).

Table 4. Robust Accuracy under Orthogonal-PGD Attack.

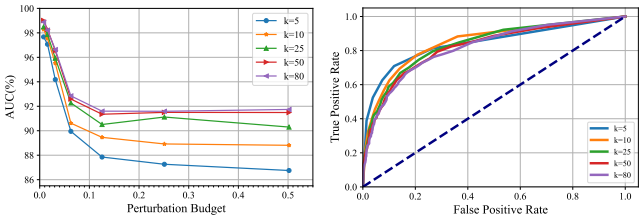
Defense	L_∞=0.01		L_∞=8/255	
	RA@FPR5%	RA@FPR50%	RA@FPR5%	RA@FPR50%
BEYOND	*88.38%*	*98.81%*	*13.80%*	*48.20%*
BEYOND +ATC	**96.30%**	**99.30%**	**94.50%**	**97.80%**
Trapdoor (Shan et al., 2020)	0.00%	7.00%	0.00%	8.00%
DLA (Sperl et al., 2020)	62.60%	83.70%	0.00%	28.20%
SID (Tian et al., 2021)	6.90%	23.40%	0.00%	1.60%
SPAM (Liu et al., 2019)	1.20%	46.00%	0.00%	38.00%

Table 5. Comparison of robust accuracy against adaptive attacks on CIFAR-10.

Classifier	Method	RA
Standard	Mao	18.97%
	BEYOND	**19.45%**
ATC	Mao	75.09%
	BEYOND	**93.20%**

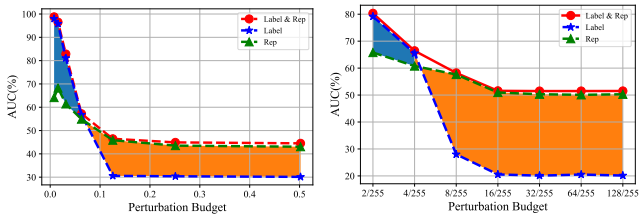
5.4. Ablation Study

The Number of Neighbors K. We examined the impact of varying the number of neighbors (K) on the detection capabilities of BEYOND against both standard and adaptive attacks, testing K values of 5, 10, 25, 50, and 80. Fig. 3 (a) illustrates how neighbor count affects performance in detecting PGD attacks across a range of perturbation budgets. We observed that detection performance generally improves with a larger neighbor set; however, gains plateau beyond K = 50. In the context of adaptive attacks, Fig. 3 (b) evaluates performance for various K values with a fixed perturbation ($\epsilon = 8/255$). Contrary to intuition, adaptive attacks are less effective with a smaller K. This is because only four linear transformations (horizontal flip, crop, color jitter, and grayscale) are deployed in BEYOND, where varying neighbors simply involve different transformation parameters. With a smaller K, the diversity among neighbors is pronounced, complicating the optimization process for adaptive attacks (multi-task learning increases model robustness (Mao et al., 2020)). Conversely, a larger K potentially results in similar neighbors that provide a wealth of information for adaptive attacks to exploit for each transformation, as detailed in the Appendix.



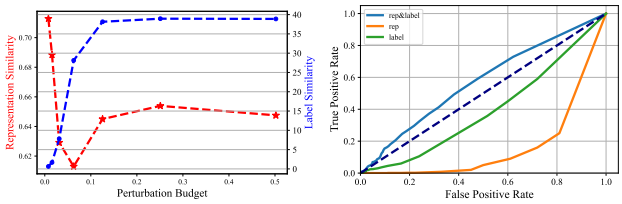
(a) Standard AEs (b) Adaptive AEs

Figure 3. Ablation Study of the Number of Neighbors.



(a) CIFAR-10 (b) IMAGENET

Figure 4. Ablation studies of representation similarity & label consistency against adaptive attacks.



(a) Representation Similarity & Label Consistency (b) Single-objective Adaptive Attacks ($\epsilon = 16/255$)

Figure 5. Trade-off between Label Consistency and Representation Similarity.

Contribution of Representation Similarity & Label Consistency against Adaptive Attacks. The analysis in Sec. 4.2 shows label consistency is more beneficial for detecting small perturbations, while representational similarity is favorable for large perturbations, which is consistent with results in Fig. 4. When the perturbation is small, the detection performance based on label consistency (blue line) is better than representation similarity (green line). As perturbation increases, representation similarity is difficult to maintain, leading to higher performance of representation similarity-based detectors. In summary, label consistency and representation similarity have different sensitivities to perturbation. Consequently, a combined approach leverages the strengths of both, culminating in superior performance (red line).

Trade-Off Between Representation Similarity and Label Consistency. The previous analysis and empirical results have proved that there is a trade-off between label consistency and representation similarity. Fig. 5 (a) shows the variation of label consistency and representation similarity with perturbation budget on CIFAR-100. It can be observed that label consistency and representation similarity respond differently to the perturbation budget, small perturbations are beneficial for representation similarity, and large perturbations favor label consistency, which matches the conclusion in Sec. 4.2. Furthermore, both objectives can be optimized simultaneously when the perturbation is large enough, which is why the adaptive attack in Fig. 2 can completely break BEYOND when the perturbation budget is larger than $16/255$. Fig. 5 (b) shows that when there is only one detection strategy, either label consistency or representation similarity, the adaptive attack can break through the defense. However, when attacking both strategies, the attack performance decreases. Hence, the robustness of BEYOND to adaptive attacks comes from the conflicts arising from optimizing these two strategies. See the Appendix for more visualization results of optimization conflicts.

Detection Performance of BEYOND Using Different SSL Models. BEYOND can flexibly cooperate with various SSL models without compromising AE detection performance, as long as the SSL model is trained to generate similar representations for the input and its augmentations. To illustrate this flexibility, we integrate BEYOND with five different SSL models: SimSiam (Chen & He, 2021) (employed in the main experiment), MoCo v3 (Chen et al., 2021), SwAV (Caron et al., 2020b), and DeepCluster v2 (Caron et al., 2020a). Table 6 presents the AE detection performance of BEYOND with these SSL models. Note that all pretrained SSL models are sourced from Solo-learn (da Costa et al., 2022). It can be seen that BEYOND demonstrates strong robustness against most adversarial attacks, consistently achieving high AUC scores (generally above 90%) across different datasets and SSL backbones. On CIFAR-10 and CIFAR-100, MoCo v3 generally yields the best results, followed closely by SimSiam and BYOL. However, since SimSiam's pretrained weights are more accessible than MoCo v3, we choose SimSiam as the backbone in this paper. In addition, MoCo v3 is a contrastive learning model that uses ViT as the backbone, while other SSL models use CNN as the backbone. the good performance of BEYOND combined with MoCo v3 shows that the performance of BEYOND does not receive the influence of the model architecture. Moreover, SwAV and DeepCluster v2 perform slightly lower than the other SSL models on CIFAR-10 and CIFAR-100. This is due to the fact that SwAv and DeepCluster v2 are clustering-based contrastive learning methods, which do not directly learn the representation similarity between input samples and their augmentations as other SSL models do, but instead learn the similarity to the clustering center, which is different

Table 6. AUC scores for BEYOND with various SSL models against adversarial attacks. SSL models trained on CIFAR-10 and CIFAR-100 are implemented with ResNet18, trained on IMAGENET are implemented with ResNet50.

Dataset	Model	FGSM	PGD	C&W	APGD-CE	APGD-T	FAB-T	Square
CIFAR-10	SimSiam	97.17%	96.48%	98.22%	96.60%	99.45%	99.14%	98.60%
	BYOL	97.22%	94.60%	98.38%	94.97%	99.54%	99.61%	99.02%
	MoCo v3	98.54%	98.26%	99.25%	98.38%	99.82%	99.69%	99.31%
	SwAV	96.29%	94.81%	97.62%	95.40%	99.14%	98.73%	98.16%
	DeepCluster v2	92.68%	89.28%	95.32%	90.72%	98.04%	97.56%	96.55%
CIFAR-100	SimSiam	97.82%	97.29%	97.93%	97.40%	98.33%	97.99%	97.80%
	BYOL	98.04%	97.00%	98.01%	96.75%	98.45%	98.33%	98.13%
	MoCo v3	98.34%	98.10%	98.50%	98.14%	98.81%	98.58%	98.44%
	SwAV	97.58%	96.91%	97.85%	97.01%	98.44%	97.94%	97.70%
IMAGENET	SimSiam	92.01%	96.88%	94.56%	97.15%	97.45%	95.47%	94.58%
	BYOL	92.01%	96.57%	94.58%	96.67%	97.00%	95.65%	94.25%

from the workflow of BEYOND. In summary, as a plug-and-play method, BEYOND can be seamlessly integrating with various SSL models.

5.5. Implementation Costs

BEYOND incorporates a supplementary SSL model for AE detection, which naturally incurs additional computational, storage and time overheads. Table 7 presents the comparison for SOTA adversarial training defense and AE detection method, i.e. LNG. The detection models have a leaner model compacity compared to ATCs, which can be reflected by the *Params* and *FLOPs* (Xie et al., 2020) being much lower than those of ATC. For BEYOND, the projection head is a three-layer FC, leading to higher parameters and *FLOPs* than LNG. However, BEYOND only compares the relationship between neighbors without calculating the distance with the reference set, resulting in a faster inference speed than that of LNG. The method of Mao et al. requires iteration, making its inference time unaffordable (Croce et al., 2022). We show the FLOPs $\times$ Params $\times$ Time as the *Overall* metric in Table 7's last column for overall comparison. If cost is a real concern in some scenarios, we can further reduce the cost with some strategy, e.g., reducing the neighbor number, without compromising performance significantly, as shown in Fig. 3 (a).

6. Conclusion

In this paper, we take the first step to detect AEs by identifying abnormal relations between AEs and their neighbors without prior knowledge of AEs. Samples that have low label consistency and representation similarity with their neighbors are detected as AE. Experiments with limited and perfect knowledge attacks show that BEYOND outperforms the SOTA AE detectors in both detection ability and efficiency. Moreover, as a plug-and-play model, BEYOND can be well integrated with ATC to further improve robustness.

Table 7. Comparison of Implementation Costs.

	Model	FLOPs(G)	Params(M)	Time(s)	Overall
ATC	(Rebuffi et al., 2021)	38.8	254.44	1.21	11945
	(Gowal et al., 2021)	38.8	254.44	1.21	11945
	(Gowal et al., 2020)	38.8	254.44	*1.21*	11945
	(Rebuffi et al., 2021)	60.57	396.23	1.24	29760
Det.	Mao	5.25	38.12	38.46	7697
	LNG	**0.286**	**8.33**	9.22	*20.521*
	BEYOND	*0.715*	*20.62*	**1.12**	**16.512**

Acknowledgements

The research work described in this paper was conducted in the JC STEM Lab of Intelligent Design Automation funded by The Hong Kong Jockey Club Charities Trust. This work is supported in part by the General Research Fund (GRF) of Hong Kong Research Grants Council (RGC) under Grant No. 14203521, in part by the CUHK SSFCRS funding No. 3136023, and in part by the Research Matching Grant Scheme under Grant No. 7106937, 8601130, and 8601440.

Impact Statement

Considering ethical concerns and the potential social effects in the future, we recommend that both users and developers adopt BEYOND as a tool to enhance robustness against adversarial attacks and adaptive attacks. We anticipate that our methodology will be particularly valuable in safety-critical scenarios.

References

Abusnaina, A., Wu, Y., Arora, S., Wang, Y., Wang, F., Yang, H., and Mohaisen, D. Adversarial example detection using latent neighborhood graph. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7687–7696, 2021.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pp. 484–501, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58591-4. doi: 10.1007/978-3-030-58592-1_29. URL https://doi.org/10.1007/978-3-030-58592-1_29.

Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., and Roundy, K. "real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 339–364. IEEE, 2023.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018a.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018b.

Bhatia, R. and Davis, C. A cauchy-schwarz inequality for operators with applications. *Linear algebra and its applications*, 223:119–129, 1995.

Bryniarski, O., Hingun, N., Pachuca, P., Wang, V., and Carlini, N. Evading adversarial example detection defenses with orthogonal projected gradient descent. *arXiv preprint arXiv:2106.15023*, 2021.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020a.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020b.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021.

Cococcioni, M., Rossi, F., Ruffaldi, E., Saponara, S., and de Dinechin, B. D. Novel arithmetics in deep neural networks signal processing for autonomous driving: Challenges and opportunities. *IEEE Signal Processing Magazine*, 38(1):97–110, 2020.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Croce, F., Gowal, S., Brunner, T., Shelhamer, E., Hein, M., and Cemgil, T. Evaluating the adversarial robustness of adaptive test-time defenses. In *International Conference on Machine Learning*, pp. 4421–4435. PMLR, 2022.

da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. solo-learn: A library of self-supervised methods for visual representation learning. *J. Mach. Learn. Res.*, 23: 56–1, 2022.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., and Roli, F. Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pp. 321–338, 2019.

Dubey, A., van der Maaten, L., Yalniz, I. Z., Li, Y., and Mahajan, D. Defense against adversarial images using web-scale nearest-neighbor search. *computer vision and pattern recognition*, 2019.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

Ho, C.-H. and Nvasconcelos, N. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.

Hu, S., Yu, T., Guo, C., Chao, W.-L., and Weinberger, K. Q. A new defense against adversarial images: Turning a weakness into a strength. *neural information processing systems*, 2019.

Jaiswal, A., Ramesh Babu, A., Zaki Zadeh, M., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning, 2020.

Jiang, Z., Chen, T., Chen, T., and Wang, Z. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.

Kaissis, G. A., Makowski, M. R., Rückert, D., and Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

Kim, M., Tack, J., and Hwang, S. J. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

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

Liu, J., Zhang, W., Zhang, Y., Hou, D., Liu, Y., Zha, H., and Yu, N. Detection based defense against adversarial examples from the steganalysis point of view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4825–4834, 2019.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., and Vondrick, C. Multitask learning strengthens adversarial robustness. In *European Conference on Computer Vision*, pp. 158–174. Springer, 2020.

Mao, C., Chiquier, M., Wang, H., Yang, J., and Vondrick, C. Adversarial attacks are reversible with natural supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 661–671, 2021.

Mikołajczyk, A. and Grochowski, M. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pp. 117–122. IEEE, 2018.

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

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Raff, E., Sylvester, J., Forsyth, S., and McLean, M. Barrage of random transforms for adversarially robust defense. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6528–6537, 2019.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Shan, S., Wenger, E., Wang, B., Li, B., Zheng, H., and Zhao, B. Y. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. *computer and communications security*, 2020.

Shi, C., Holtz, C., and Mishne, G. Online adversarial purification based on self-supervision. *arXiv preprint arXiv:2101.09387*, 2021.

Sperl, P., Kao, C.-Y., Chen, P., Lei, X., and Böttinger, K. Dla: dense-layer-analysis for adversarial example detection. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 198–215. IEEE, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tian, J., Zhou, J., Li, Y., and Duan, J. Detecting adversarial examples from sensitivity inconsistency of spatial-transform domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9877–9885, 2021.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–828, 2020.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Yang, Y., Gao, R., Li, Y., Lai, Q., and Xu, Q. What you see is not what the network infers: Detecting adversarial examples based on semantic contradiction. In *Network and Distributed System Security Symposium (NDSS)*, 2022.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Zeng, Y., Qiu, H., Memmi, G., and Qiu, M. A data augmentation-based defense method against adversarial attacks in neural networks. In *International Conference on Algorithms and Architectures for Parallel Processing*, pp. 274–289. Springer, 2020.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zuo, F. and Zeng, Q. Exploiting the sensitivity of l2 adversarial examples to erase-and-restore. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pp. 40–51, 2021.

A. Datasets & Models

We conduct experiments on three commonly adopted datasets including CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100, and a more IMAGENET (Krizhevsky et al., 2012). The details of the target models (classifiers), and the employed SSL models together with their original classification accuracy on clean samples are summarized in Table 8 [5].

Table 8. Information of datasets and models.

Dataset	Classifier SSL	Acc. on clean samples↑	
		Classifier	SSL
CIFAR-10	ResNet18	91.53%	90.74%
CIFAR-100	ResNet18	75.34%	66.04%
IMAGENET	ResNet50	80.86%	68.30%

B. Detection Performance

B.1. TPR@FPR against Limited Knowledge Attacks.

Table 9 reports TPR@FPR5% to show the AE detection performance of BEYOND. It can be observed that BEYOND maintains a high detection performance on various attacks and datasets, which is attributed to our detection mechanism. Combining label consistency and representation similarity, BEYOND identifies AEs without reference AE set.

Table 9. TPR@FPR 5% of BEYOND against Limited Knowledge Attacks. All attacks are performed under $L_\infty = 8/255$.

Dataset	CIFAR-10	CIFAR-100	IMAGENET
Attack	**TPR@FPR5% ↑**		
FGSM	86.16%	89.80%	61.05%
PGD	82.80%	85.90%	89.80%
C&W	91.48%	91.96%	76.69%
AutoAttack	93.42%	90.90%	84.25%

Table 10 reports TPR@FPR 3% to further demonstrate the AE detection capability of BEYOND. Because the detection mechanism does not rely on additional prior knowledge of AE or model retraining, it has been confirmed that BEYOND can generalize well to defend various attacks. Furthermore, on the complex dataset, i.e., IMAGENET, BEYOND still maintains a high detection performance.

Table 10. TPR@FPR 3% of BEYOND against Limited Knowledge Attacks. All attacks are performed under $L_\infty = 8/255$.

Dataset	CIFAR-10	CIFAR-100	IMAGENET
Attack	**TPR@FPR3% ↑**		
FGSM	76.37%	81.93%	51.74%
PGD	69.50%	76.00%	82.20%
C&W	85.29%	84.32%	68.50%
AutoAttack	88.33%	83.91%	72.06%

B.2. Accuracy with ATC

Following (Yang et al., 2022), Accuracy in Table 3 indicates the detector's accuracy on clean samples by combining the detector with the classifier, and calculated as follows:

[5]The pre-trained SSL models for CIFAR-10 and CIFAR-100 are from Solo-learn (da Costa et al., 2022), and for IMAGENET are from SimSiam (Chen & He, 2021).

$$Acc = \frac{\#\text{Classifier correct\&Detector pass}}{\#\text{All clean samples}} + \frac{\#\text{Classifier wrong\&Detector reject}}{\#\text{All clean samples}}$$

B.3. Performance against Orthogonal-PGD Selection Strategy Adaptive Attacks

Orthogonal-Projected Gradient Descent (Orthogonal-PGD) is a recently proposed AE detection benchmark. In the selection strategy, Orthogonal-PGD updates the input by selectively exploiting perturbations produced by either the classifier or the detector to avoid perturbation waste. Table 11 shows BEYOND outperforms the four baselines by a considerable margin in selection strategy, especially under small perturbations.

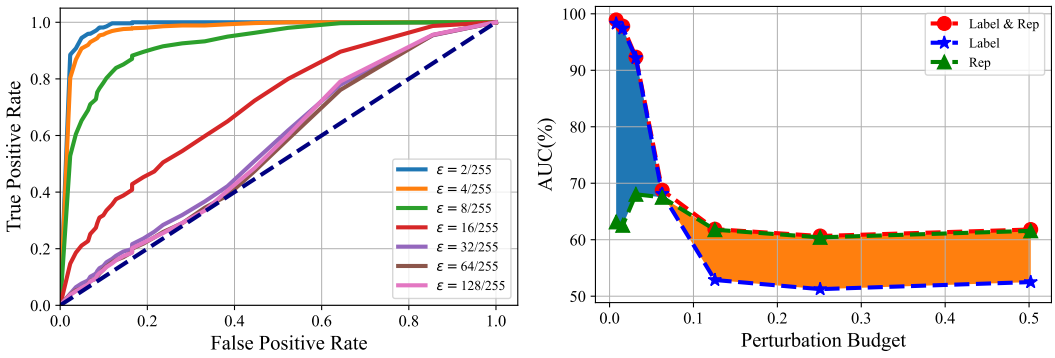
For the worst case, BEYOND can still maintain 8.04% ($L_\infty = 8/255$), while the baselines are only 0.4%. Furthermore, incorporating ATC can significantly improve the detection performance of BEYOND against large perturbation to 91.5%.

Table 11. Robust Accuracy under Orthogonal-PGD selection strategy on CIFAR-10. The **bolded** values are the best performance and the *underlined italicized* values are the second-best performance.

Defense	L_∞=0.01		L_∞=8/255	
	RA @FPR5%	RA @FPR50%	RA @FPR5%	RA @FPR50%
BEYOND	*79.63%*	*97.47%*	*8.04%*	*40.42%*
BEYOND +ATC	**95.80%**	**99.40%**	**91.50%**	**95.90%**
Trapdoor	0.20%	49.50%	0.40%	37.20%
DLA'20	17.00%	55.90%	0.00%	13.50%
SID'21	8.90%	50.90%	0.00%	11.40%

B.4. Detection Performance on CIFAR-100

Fig. 6 shows the detection performance of BEYOND against adaptive attacks on CIFAR-100 and the contribution of label consistency and representation similarity. It can be seen BEYOND is effective for detecting adaptive attacks on CIFAR-10. Meanwhile, label consistency is more suitable for detecting small perturbations, while representation similarity is favourable for large perturbations, which is consistent with the conclusion on CIFAR-10 and IMAGENET.



(a) (b)

Figure 6. (a) Detection performance against adaptive attacks on CIFAR-100. (b) Contribution of label consistency and representation similarity on CIFAR-100

B.5. Detection Performance for Various Types of Attacks

To evaluate the detection performance of BEYOND for different types of attacks, we test the most representative method that supports multiple norm attacks, AutoAttack. AutoAttack supports L_∞, L_2 and L_1 norm attacks. In the main paper, we only

Table 12. AUC for Adaptive Attack under different α.

α	**0**	**1**	**10**	**20**	**50**	**100**
CIFAR-10	82.03%	63.91%	64.57%	76.15%	88.56%	92.53%
CIFAR-100	90.58%	88.49%	91.61%	93.10%	94.05%	94.37%

Table 13. Detection performance of BEYOND against AutoAttack with different norms.

AUC(%)	L_∞	L_2	L_1
CIFAR-10	99.18	99.13	99.07
IMAGENET	97.14	97.26	97.18

report the detection performance of BEYOND against AutoAttack L_∞. Table 13 shows the performance of BEYOND against AutoAttack with different norms. Where the perturbation budgets (ϵ) on CIFAR-10 are $8/255$ (L_∞), 0.5 (L_2), and 8 (L_1); and on IMAGENET are $8/255$ (L_∞), 3 ($L2$), and 64 (L_1). The results show BEYOND is still effective against attacks based on different norms.

B.6. Hyperparameter Alpha in Adaptive Attacks

The design of the adaptive attack in Eq. 11 includes a hyperparameter α, which is a trade-off parameter between label consistency and representation similarity. Table 12 shows the AUC of BEYOND under different α. As shown, when $\alpha = 0$, i.e. the attacker only attacks the label consistency detection mechanism, the AUC score is still high, which proves that our approach is not based on the weak transferability of AEs. Moreover, adaptive attacks are strongest when $\alpha = 1$, which is used for all tests.

Table 14. Performance Comparison Using Cutmix and Mixup.

AUC (%)	α	FGSM	PGD	CW	AutoAttack	Avg
Cutmix	1.0	93.69	94.96	96.10	94.69	94.86
	0.7	93.87	95.28	96.33	94.70	95.05
	0.5	94.15	94.69	96.75	95.33	**95.23**
Mixup	1.0	89.03	89.07	89.32	89.20	89.16
	0.7	89.43	89.36	89.87	89.60	89.57
	0.5	90.15	89.35	90.01	90.03	**89.89**
BEYOND	-	98.89	99.28	99.20	99.16	**99.13**

B.7. Detection Performance Using Cutmix & Mixup

BEYOND employs a set of augmentations—horizontal flipping, cropping, color jitter, and grayscale—to generate neighbors. These augmentations are aligned with those used for training the SSL model. Thanks to the feature invariance of SSL models to input transformations, BEYOND can efficiently detect adversarial samples without compromising the accuracy of benign samples. Therefore, we think that augmentation methods with minimal effect on image representation are more advantageous for detection, as aggressive augmentations could significantly alter the features of benign samples.

We tested the detection performance for BEYOND using Cutmix (Yun et al., 2019) and Mixup (Zhang et al., 2017) as augmentation methods on CIFAR-10 in Table 14. It should be noted that we employ Cutmix and Mixup from the torchvision library, which includes an α hyperparameter to control the mix ratio. Observations are as follows: a) Cutmix and Mixup are not as effective as BEYOND because they merge two images, which has a more substantial impact on the feature representation compared to standard data augmentation methods. b) Cutmix has an edge over Mixup since Cutmix only integrates a portion of one image into another, whereas Mixup combines the entirety of both images. Hence, Cutmix has a less pronounced effect on the image features, leading to its superior performance. c) A reduction in the α value diminishes the mixing's influence on the image features, which in turn enhances the performance of both Cutmix and Mixup. This improvement aligns with our expectations.

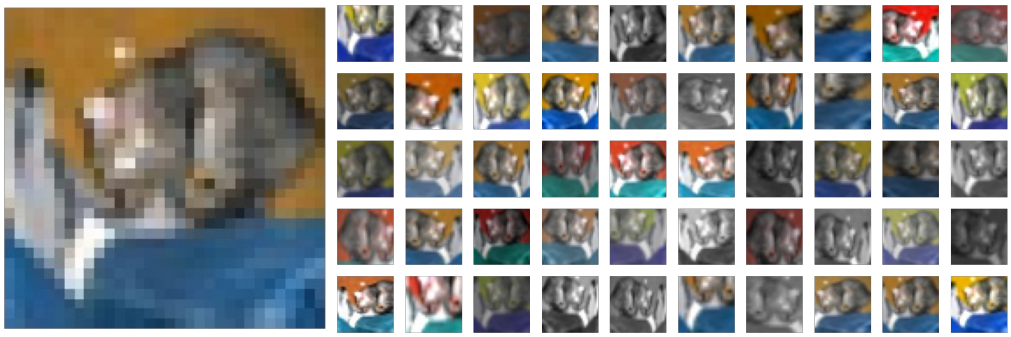


Figure 7. Display of generated neighbors. The original image is on the left and the generated 50 neighbors are on the right.

C. Display of Generated Neighbors

Fig. 7 shows the 50 neighbors augmented by the original image. Augmentations are made up of four linear variations including color jitter, crop, horizontal flip and greyscale. Neighbors are generated by random combinations of transformation parameters, whose consistency is ensured by fixing random seeds. It can be noticed that when the number of generated neighbors is small, there is a large difference between neighbors, while when the number of generated neighbors is large, there are similar neighbors. This may be the reason why the adaptive attack is a little more difficult to break BEYOND when k is small in Fig. 3.

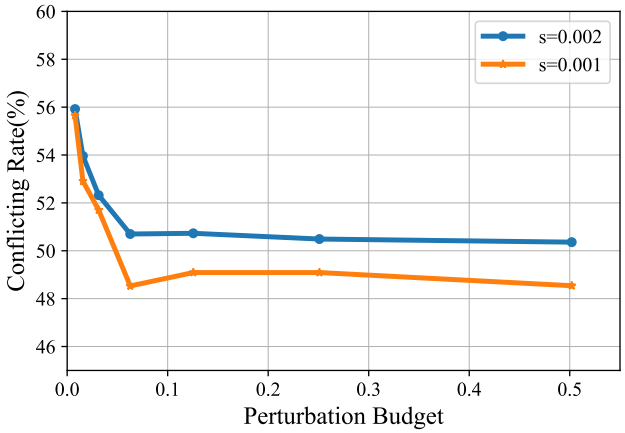


Figure 8. Conflicting rate for optimizing label consistency and representation similarity with different attack step sizes.

Table 15. The ratio of different data augmentations meeting the threshold. Compose is a combination of augmentations used to train SSL models, including crop, resize, horizontal flip, and color jitter.

Aug	Ratio	Aug	Ratio
Rotation	99.9%	Vertical	25.9%
Crop	40.7%	Color Jitter	99.0%
Resize	74.0%	Gray	40.6%
Horizontal	25.9%	Compose	99.7%

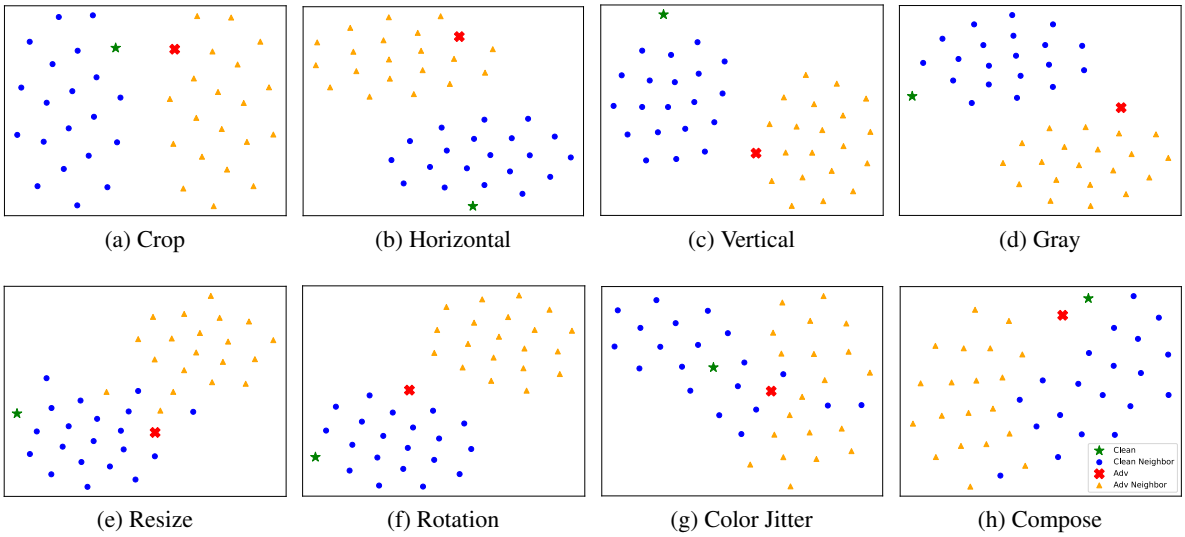
D. Select Effective Augmentations

To better improve the effectiveness of BEYOND, we analyze the conditions under which the augmentation can effectively weaken adversarial perturbation. Effective data augmentation makes the augmented label y_{aug} tend to the ground-truth label y_{true} and away from the adversarial label y_{adv}:

$$||y_{aug} - y_{true}||_2 \leq ||y_{aug} - y_{adv}||_2 \leq ||y_{adv} - y_{true}||_2. \tag{16}$$

Since y_{true} is the one-hot encoding, the range of $||y_{adv} - y_{true}||_2$ is $(\sqrt{2}/2, \sqrt{2})$. The distance is $\sqrt{2}$ when the item corresponding to y_{adv} is 1 in the logits of y_{adv}, and $\sqrt{2}/2$ when the item corresponding to y_{adv} and y_{true} both occupy 1/2. Given a SSL-based classifier, C, we have:

$$\begin{aligned} C(W(x+\delta)) &= C(Wx) + \nabla C(Wx)W\delta \\ &= y_{true} + \nabla C(Wx)W\delta = y_{aug}. \end{aligned} \tag{17}$$



(a) Crop (b) Horizontal (c) Vertical (d) Gray

(e) Resize (f) Rotation (g) Color Jitter (h) Compose

Figure 9. Visualization of clean sample and AE with different augmented neighborhoods.

Therefore, the distance between y_{aug} and y_{true} is:

$$\begin{aligned} &||y_{aug} - y_{true}||_2 = ||\nabla C(Wx)W\delta||_2 \\ &\leq ||\nabla C(Wx)W||_2||\delta||_2 \leq ||\nabla C(Wx)W||_2\epsilon \end{aligned} \tag{18}$$

where $||\delta||_2$ is bounded by ϵ. Eq. 16 always holds, then:

$$||\nabla C(Wx)W||_2\epsilon \leq \frac{\sqrt{2}}{2} \Rightarrow ||\nabla C(Wx)W||_2 \leq \frac{\sqrt{2}}{2\epsilon}. \tag{19}$$

In summary, augmentation can mitigate adversarial perturbation when it satisfies Eq. 19.

To further validate our analysis, we generate 1000 adversarial examples by PGD with $\epsilon = 8/255$ on CIFAR-10. Table 15 shows the ratios for different data augmentations meeting the threshold $\frac{\sqrt{2}}{2\epsilon}$. A higher ratio means the augmentation is more effective. It can be observed that Rotation, Color Jitter and Compose are the three most effective augmentations according to our analysis. To further validate our analysis, we perform t-sne (Van der Maaten & Hinton, 2008) visualizations of the SSL representations of clean and AEs processed by different augmentation methods. We utilize a self-supervised feature extractor and projection head to obtain SSL representations and use augmentation methods to generate 20 neighbors for both clean samples and AEs. As seen in Fig. 9, the effective augmentation methods with the high ratio in Table 15 can effectively increase the distance between AEs and their neighbors. For example, Rotation has the highest ratio in Table 15, and the distance between AE and its neighbors in Fig. 9 is larger than that of clean samples. While Horizontal and Vertical have the lowest ratio, and the distance between AE and its neighbors is still close in Fig. 9

Table 16. Detection performance comparison of augmentations.

Augmentation	FGSM	PGD	CW	AutoAttack	Average
ColorJitter&Resize&Rotation	97.11%	96.55%	98.15%	96.56%	97.09%
Gray&Horizaotal&Crop&Vertical	92.44%	91.36%	94.70%	91.87%	92.59%

Moreover, we test the detection performance of high-ratio augmentations and low-ratio augmentations in Table 16. It can be seen that the average detection performance of the effective augmentations obtained by our analysis is 5% higher than that of the other augmentations.

E. Conflict Rate of Label Consistency and Representation Similarity

The conflict between label consistency and representation similarity stems from their different optimization goal. Fig. 8 shows the gradient conflict rate for adaptive attacks with different step sizes on different perturbation budgets. We can find that the gradient conflict rate decreases for large perturbations and converges as the perturbation further increases, with the convergence point being consistent with the turning point in Fig. 5 of the main paper.

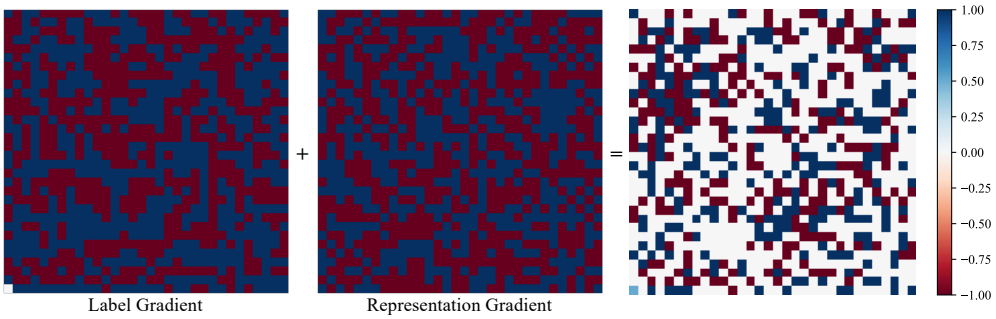


Figure 10. Gradient conflict between label consistency and representation similarity. The colored pixels represent the gradient direction, while the blank means gradient conflict.

Sec. 4.2 demonstrates the conflict between label consistency and representation similarity stems from their different optimization goals. Fig. 10 visualizes the gradients produced by optimizing label consistency and representation similarity on the input. It's shown that attacks on label consistency or representation similarity produce gradients that modify the input in a certain direction, but optimizing for both leads to conflicting gradients. The experiments in Fig. 8 show that the gradient conflict rate decreases when the perturbation becomes larger, which is consistent with the results in Fig. 5 (a).