A Simple Class-Agnostic Approach to Enhance Fair Adversarial Training

Erh-Chung Chen[1]
s107062802@m107.nthu.edu.tw

Pin-Yu Chen[2]
pin-yu.chen@ibm.com

I-Hsin Chung[2]
ihchung@us.ibm.com

Che-Rung Lee[1]
cherung@cs.nthu.edu.tw

[1]**National Tsing Hua University**
[2]IBM Research

Abstract

AI security has become a critical issue as Deep Neural Networks (DNNs) are inherently vulnerable to adversarial attacks. While previous studies have demonstrated that adversarial training is one of the most effective defensive strategies, improvements in robustness often show significant class-wise disparities. This has established **fair adversarial training**, which aims to enhance the robustness of the worst-performing classes, as a crucial research direction. Although existing class-wise approaches can improve performance by adjusting individual class weights, they face severs scalability limitation on large-scale datasets due to the sparse volume of data available per class. To mitigate this issue without relying on explicit class-wise information, we propose a novel **tri-regime optimization framework** that fundamentally decouples the distinct sources of adversarial error at the sample level. Specifically, our method systematically suppresses unlearnable outliers via reliability gating, prioritizes genuine boundary threats through dynamic reweighting, and down-weights already *safe* examples to prevent redundant optimization. By isolating and prioritizing the effective adversarial frontier, our method strictly prevents robust overfitting and achieves superior worst-class robustness while maintaining high clean accuracy. Furthermore, we demonstrate that applying solely the regularization mechanism serves as a highly efficient post-training strategy on massive datasets like ImageNet while maintaining high clean accuracy.

1 Introduction

Deep Neural Networks (DNNs) have demonstrated remarkable performance across various fields [1, 2]. These advances have catalyzed the development of numerous applications now embedded in our daily lives, from chatbots [3] to self-driving cars [4]. However, despite their remarkable capabilities, DNN-based AI applications face significant security challenges that remain inadequately addressed. A major threat among these is adversarial attacks, where subtle, imperceptible perturbations are introduced into input data, causing AI models to make highly confident but completely incorrect predictions. These vulnerabilities pose severe risks, particularly in safety-critical domains such as autonomous driving systems [5], unmanned aerial vehicles (UAVs) [6], and surveillance systems [7]. These extreme failures are not merely edge cases; they represent real and present dangers to human safety, highlighting the fragility of AI systems.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

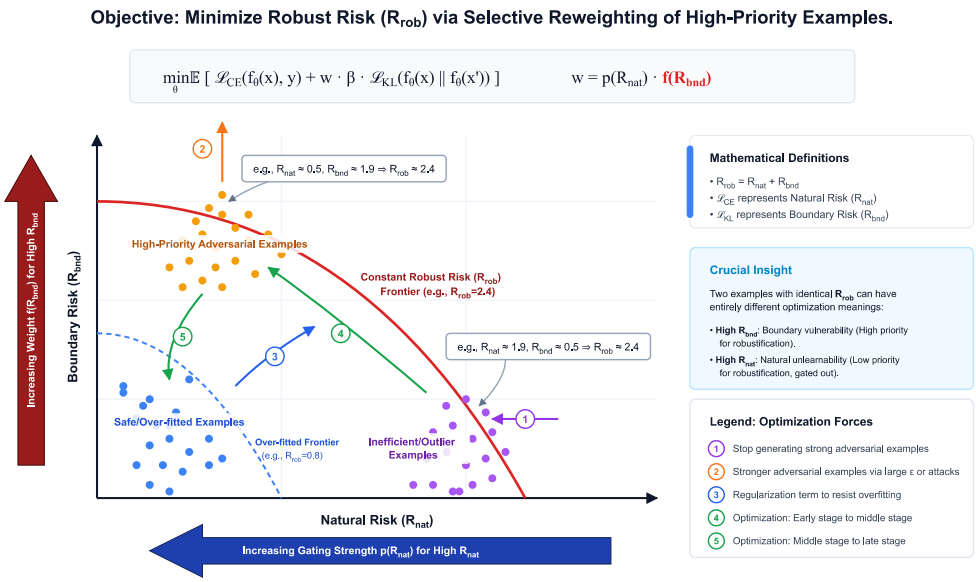


Figure 1: **Conceptual overview of the proposed Tri-Regime Selective Reweighting framework.** Examples lying on the same *Constant Robust Risk Frontier* (red curve, e.g., $R_{rob} \approx 2.4$) can have entirely different optimization implications based on their composition of Natural Risk (R_{nat}) versus Boundary Risk (R_{bnd}). Standard methods treat the "Inefficient/Outlier Examples" (purple cluster) exactly the same as "High-Priority Adversarial Examples" (orange cluster), inevitably leading to robust overfitting on noise. To resolve this ambiguity, our framework explicitly decouples these error sources using three tailored mechanisms: (1) a **Reliability Gating** term $\rho(R_{nat})$ (bottom blue arrow) that suppresses high-noise outliers; (2) a **Dynamic Reweighting** term $\phi(R_{bnd})$ (left red arrow) that prioritizes genuinely vulnerable boundary threats; and (3) a **Sample-Level Regularization** mechanism that diminishes the weight of "Safe/Over-fitted" examples (blue cluster) to prevent wasted capacity. This targeted strategy successfully steers the optimization trajectory (green arrows) toward the robust frontier while strictly avoiding the memorization of unlearnable data.

Adversarial training [8, 9, 10], one of the most extensively studied and effective defense mechanisms against adversarial attacks, fundamentally relies on augmenting the training data with adversarial examples and optimizing the model to minimize the losses caused by these perturbations. Standard approaches treat these examples uniformly; however, certain natural examples are inherently unlearnable or fail to yield sufficiently strong adversarial perturbations. This indiscriminate optimization forces the model to memorize meaningless noise, which inevitably causes robust overfitting [11] and ultimately leads to severe class-wise disparities in robustness.

To address this disparity, the sub-field of **fair adversarial training** has emerged, aiming to boost the robustness of the worst-performing classes. However, existing fair approaches typically rely on coarse reweighting strategies that indiscriminately prioritize vulnerable classes based on the overall robust risks of adversarial examples belonging to those classes. While these class-wise strategies offer improvements, it remains fundamentally flawed because it fails to recognize instance-level (intra-class) variance, treating all data within a vulnerable class as equally important. Additionally, these strategies typically require splitting off a separate validation set to compute dynamic class statistics. This imposes severe scalability limitations on large-scale datasets, as the per-class data volume available for reliable validation becomes prohibitively sparse.

In this paper, **we argue that class-wise information is actually unnecessary; the key to robust fairness is dynamically assigning optimization priority at the individual sample level based on its specific loss composition.** As illustrated in Figure 1, examples with same robust risk (red curve) can possess entirely different properties. To resolve this ambiguity, we propose a novel **tri-regime optimization framework** that explicitly decouples the distinct sources of adversarial

risks at the sample level, replacing standard uniform weighting with three tailored mechanisms: First, examples in the "Inefficient/Outlier" cluster (purple) dominate their loss via high natural risk, indicating ill-defined adversarial examples. We apply a gating mechanism to filter these out, halting the generation of strong attacks or assigning them near-zero weight (action ①). Second, examples in the "High-Priority" cluster (orange) exhibit high boundary risk, representing genuine adversarial threats. We employ a dynamic reweighting strategy to concentrate model capacity here, either generating stronger attacks (action ②) or adaptively up-weighting their importance. Third, examples falling within the "Safe/Over-fitted" region (blue) are correctly classified and far from the decision boundary. Continuing to optimize them wastes capacity; thus, we introduce a regularization mechanism to dynamically down-weight these examples and prevent redundant optimization (action ③). Ultimately, this targeted, three-tiered approach successfully steers the overall optimization trajectory (green arrows ④ and ⑤) toward the robust frontier while strictly avoiding the pitfall of robust overfitting.

In summary, the main contributions of this paper are as follows:

- We identify that severe class-wise disparities in robustness are the direct result of robust overfitting, driven by standard approaches treating all adversarial examples uniformly regardless of their loss composition.
- We propose a novel tri-regime optimization framework featuring reliability gating (to suppress outliers), dynamic reweighting (to prioritize genuine boundary threats), and regularization (to down-weight safe examples) entirely at the sample level, removing the need for coarse class-wise priors.
- The self-comparison enables an adaptive adjustment based on the behavior of individual samples, without relying on class-wise statistics or a separate validation set.
- We develop a highly efficient post-training extension of our approach and empirically demonstrate that, for any adversarial training model, applying only the sample-level regularization is sufficient to mitigate robust overfitting and unlock fair robustness. Notably, this enables our method to successfully scale to the massive ImageNet dataset, a milestone unaccomplished by existing class-wise works.

2 Related Work

2.1 Adversarial Training

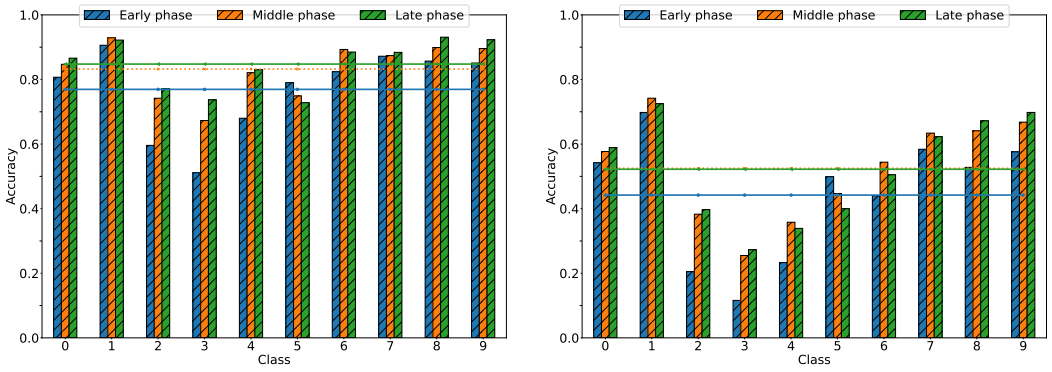
Adversarial training [8] is one of the most widely studied and effective defense strategies against adversarial attacks—malicious inputs crafted by applying tiny, often imperceptible perturbations to input data in order to mislead the victim models. The goal of adversarial training is to enhance model robustness by explicitly incorporating these adversarial examples into the training process. Formally, adversarial training can be formulated as a min-max optimization problem:

$$\min_{\theta} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \max_{\tilde{x}:D(\tilde{x},x)\leq\epsilon} \mathcal{L}(Z(\tilde{x};\theta),y), \tag{1}$$

where x and y represent the input and corresponding label; $\tilde{x}$ is the adversarial example; $D(\cdot)$ denotes a distance metric; $Z(\cdot)$ is the model with parameters θ; and $\mathcal{L}$ is the loss function. The inner maximization seeks the worst-case perturbation within an ϵ-ball around the input while the outer minimization aims to minimize the worst-case loss within the allowed perturbation region.

Over time, numerous variations of adversarial training have been developed. Some approaches aim to enhance robustness by incorporating additional training samples [10], while others focus on reducing computational cost by simplifying adversarial example generation or filtering out less informative samples [12].

Sample-wise selection and reweighting. Beyond data filtering, GAIRAT [13] assigns instance weights based on the number of attack steps needed for misclassification, while MART [14] weights its KL regularization more heavily for natural examples with low ground-truth confidence. Our boundary-loss weighting instead combines a natural correctness/confidence gate with boundary risk to prioritize vulnerable but naturally learnable examples.



(a) Natural accuracy (b) Robust accuracy

Figure 2: Evolution of natural and robust accuracy of an adversarially trained model across different training phases, using the commonly adopted TRADES framework [9] on the CIFAR-10 dataset. Horizontal lines indicate the average accuracy across all classes.

2.2 Robust Fairness

Robust fairness is an emerging topic in adversarial machine learning that addresses the imbalance in robustness across different classes. It is often observed that while some classes achieve substantial improvements in robustness, the worst-performing classes remain highly vulnerable even after adversarial training. This disparity poses a significant challenge in safety-critical applications, where consistent and certifiable robustness across all classes and edge cases is essential to ensure system reliability and safety.

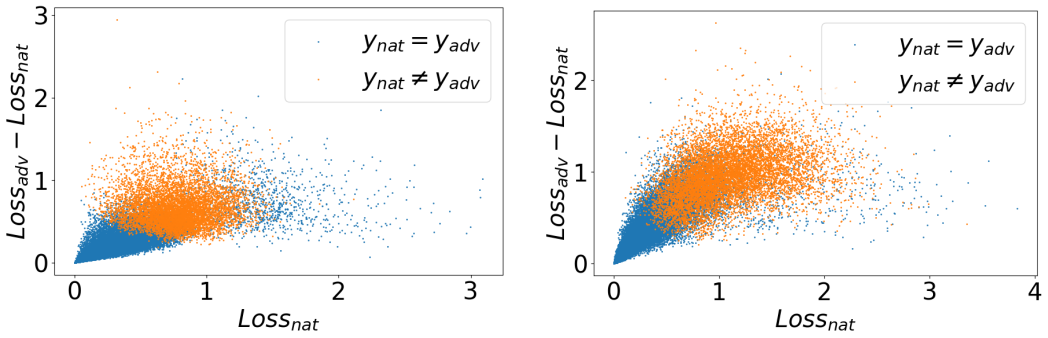
The issue of robust fairness was first explicitly addressed by FRL [15], which introduced a re-weighting policy to improve the robustness of the weakest classes. Following this, several approaches have been proposed that rely on class-wise robustness evaluation. WAT [16] adaptively adjusted class-wise weights at each training epoch based on historical performance on the validation set. Similarly, CFA [17] adopted a related strategy that not only reweights adversarial examples but also dynamically adjusts the radius of the ϵ-ball for each class, applying stronger adversarial perturbations to classes with higher robustness. BAT [18] modified the adversarial example generation process and updates model parameters with a focus on samples near decision boundaries. In contrast, FAAL [19] proposed an on-the-fly approach that mitigates class distributional shifts using a Distributionally Robust Optimization (DRO) framework. Despite their effectiveness, these methods share several limitations. Most require a validation set split from the training data, which reduces the volume of data available for model optimization and may lead to suboptimal final performance.

3 Methodology

In this section, we first conduct a pilot study to better understand the phenomenon of robustness imbalance across different classes. We argue that the root cause is robust overfitting. Reducing the weights of overfitted samples or filtering them out is equivalent to increasing the relative importance of highly vulnerable adversarial examples. Therefore, **we propose an optimization that can adjust weight at the individual sample level without requiring a separate validation set or any class-specific information.**

3.1 Pilot Study

To better understand the origins of the fairness issue, we conducted a pilot study that used the widely adopted TRADES [9] framework to analyze the evolution of natural accuracy and robust accuracy on the CIFAR-10 dataset. As shown in Figure 2, we observe three key phenomena: (1) natural and robust accuracy trends are highly correlated; (2) a substantial, persistent performance gap exists between the best- and worst-performing classes; and (3) the relative hierarchy of class rankings and average accuracy stabilizes remain rigidly fixed throughout the middle and late training phases. This



(a) CIFAR-10 dataset.

(b) CIFAR-100 dataset.

Figure 3: Distribution of cross-entropy losses for natural examples and their corresponding adversarial examples, generated using a fixed-step attack over the training set. y_{nat} and y_{adv} denote the predicted labels of natural data and the corresponding adversarial examples, respectively.

static behavior dictates that worst-performing classes almost never recover to join the top-performing group, a phenomenon consistently observed across various adversarial training frameworks.

Consequently, while dynamic class-wise re-weighting presents an intuitive countermeasure, existing methodologies remain fundamentally constrained. Because the performance discrepancy emerges early and remains rigid, continuously tracking historical training dynamics as proposed by methods like CFA yields marginal empirical benefits and unnecessarily incurs the overhead of maintaining dynamic validation splits. More importantly, coarse class-wise re-weighting completely overlooks instance-level (intra-class) variance. Even within highly robust classes, there are still specific adversarial examples that exhibit high losses or remain severely under-fitted.

We argue that the fundamental limitation of these existing re-weighting strategies is inherently tied to robust overfitting. As illustrated previously in Figure 1, adversarial examples exhibiting the exact same robust risk can possess entirely different optimization properties. While genuine boundary threats inherently demand high optimization priority, they risk being drowned out by the sheer volume of "Safe/Over-fitted" examples, which heavily dominate the overall population, particularly in the late training phase. Because these over-fitted adversarial examples are already correctly classified and far from the decision boundary, they definitively deserve lower importance. However, the outsized influence of this specific group on robust overfitting has remained largely unstudied in existing works.

To provide empirical evidence for this phenomenon, Figure 3a illustrates the empirical distribution of natural risk (x-axis) versus boundary risk (y-axis), generated using a fixed-step attack over the training set. To make the two losses in the same scale, the natural risk and the boundary risk are set to cross-entropy, the difference between them. Strikingly, this empirical distribution closely mirrors the theoretical clusters proposed in Figure 1. The dense cluster of blue data points ($y_{nat} = y_{adv}$) in the bottom-left corner represents the "Safe/Over-fitted" examples, which constitute a massive fraction of the training population. As the optimization progresses, this over-fitted population gradually expands. Because these samples typically show only minimal increases in loss and their predicted labels remain unchanged under attack, continuing to uniformly optimize them wastes critical model capacity. Consequently, to successfully mitigate robust overfitting, it is imperative to shift away from coarse class-wise tracking. Instead, a framework must isolate optimization priority at the individual sample level, deliberately assigning higher weight to the vulnerable orange data points ($y_{nat} \neq y_{adv}$) while effectively diminishing the blue ones. Furthermore, as shown in Figure 3b, the loss distribution on CIFAR-100 exhibits an almost identical structural pattern, indicating that this population dynamic is a general phenomenon underlying robust overfitting across different datasets.

3.2 Class-agnostic Training Framework

According to the terminology defined by TRADES [9], the robust risk (R_{rob}) of any example can be decomposed into two distinct sources: natural risk (R_{nat}), representing inherent unlearnability, and boundary risk (R_{bnd}), representing decision boundary vulnerability:

$$R_{rob} \leq R_{nat} + \beta R_{bnd}, \tag{2}$$

where β is a positive balancing hyperparameter. To mitigate robust overfitting without relying on coarse class-wise statistics, we propose a class-agnostic training framework that dynamically adjusts optimization priority entirely at the sample level. The objective function of our proposed framework is defined as:

$$\mathcal{L}(x, x', y) = \mathcal{L}_{\text{CE}}(x, y; \theta) + \beta\omega(x, x', y)\mathcal{L}_{\text{KLD}}(x, x'; \theta), \tag{3}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy loss on the natural sample x; $\mathcal{L}_{\text{KLD}}$ represents the Kullback–Leibler (KL) divergence between the model outputs on the natural and adversarial samples; and $\omega(x, x', y) \in [0, 1]$ is our proposed sample-level masking function. The $\mathcal{L}_{\text{CE}}$ term promotes accuracy on natural examples, while the $\mathcal{L}_{\text{KLD}}$ term forces the adversarial prediction to match the natural prediction, thereby enforcing boundary robustness. Notably, if $\omega(x, x', y)$ is set to a constant value of 1, the objective in Equation (3) elegantly degenerates to the standard TRADES framework.

To operationalize our tri-regime framework, the masking function ω explicitly disentangles the treatment of outliers, genuine threats, and over-fitted data by multiplying two distinct terms: a reliability gate $\rho(x, y)$ and a dynamic reweighting factor $\phi(x, x')$, such that

$$\omega(x, x', y) = \rho(x, y) \cdot \phi(x, x'). \tag{4}$$

The joint **Reliability Gating and Regularization** function, $\rho(x, y)$, is defined as:

$$\rho(x, y) = \begin{cases} 0 & \text{if } \arg\max Z(x) \neq y \\ p_y & \text{otherwise,} \end{cases} \tag{5}$$

where p_y is the predicted probability of the ground truth class y on the natural example. This mechanism successfully identifies and masks out the dense cluster of "Safe/Over-fitted" examples residing in the bottom-left corner of the loss distribution. For the remaining active examples, the assigned weight p_y fulfills our *Reliability Gating* regime. Because p_y scales inversely with natural risk, inherently noisy outliers (which exhibit high R_{nat}) yield low p_y scores, gracefully suppressing their influence on the gradient.

The primary goal of the **Dynamic Boundary Reweighting** algorithm, $\phi(x, x')$, is to dynamically adjust the importance of adversarial examples based on their individual loss. However, the absolute scale of the loss can vary significantly across different datasets, a discrepancy that severely impacts optimization dynamics, particularly during the early training stages. To systematically tackle this scale issue, our framework offers a solution that normalizes the loss scale into the $[0, 1]$ interval. While the specific activation function can be customized, we propose the following formulation to effectively compress the input and prioritize genuine boundary threats:

$$\phi(x, x') = \tanh\left(\gamma \frac{l_{x,x'}}{l^{(t)}_{\text{anchor}}}\right), \tag{6}$$

where $l_{x,x'}$ is the KL divergence loss of the sample, and $l^{(t)}_{\text{anchor}}$ acts as the normalization anchor. To prevent skewing from extreme outliers, this anchor is explicitly set to the 98-th percentile of the KL loss across the dataset. The hyperparameter γ (typically set between 3 and 5) shapes the slope of the activation; a higher γ steepens the curve toward a step function, aggressively up-weighting high-priority adversarial examples to maximize optimization pressure where the boundary is most vulnerable.

On the other hand, because the overall scale of the loss inherently drifts and decreases as training progresses, an additional mechanism is required to dynamically adjust this anchor. To ensure the anchor accurately reflects the shifting loss landscape and to prevent stale values, we update it periodically using an Exponential Moving Average (EMA):

$$l^{(t)}_{\text{anchor}} = \tau \cdot l^{(t-1)}_{\text{anchor}} + (1 - \tau) \cdot l_{\text{anchor}}. \tag{7}$$

Finally, to operationalize the **Sample-Level Regularization** mechanism targeting the "Safe/Over-fitted" examples, we address the target distribution itself. Because ground-truth labels are conventionally encoded as one-hot vectors, standard optimization inevitably drives the model toward absolute overconfidence. To explicitly prevent this saturation, we dynamically reshape the target distribution by feeding the model's current output probabilities as the input concentration parameters for a Dirichlet distribution. Resampling from this parameterized Dirichlet distribution generates smoothed, continuous soft targets. This elegantly penalizes overconfidence on already "safe" examples, providing the vital regularization force needed to prevent capacity waste and maintain robust generalization.

Table 1: Reproduced standard and robust accuracy on CIFAR-10 dataset using WRN-34-10 architecture. The row marked with a dagger (†) indicates that a post-trained strategy was adopted

Method	Average		Worst	
	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}
TRADES	82.02	51.32	64.30	21.60
CFA	71.17	45.27	53.10	28.10
WAT	85.05	44.90	73.40	23.60
AWP	84.69	56.46	68.00	25.30
Ours	84.90	49.65	73.30	24.10
Ours+AWP	85.64	55.54	71.70	28.60
AWP†	84.57	54.95	71.50	29.00

Table 2: Standard and robust accuracy on CIFAR-100 dataset using ResNet-18 architecture. The row marked with an asterisk (*) indicates additional configurations applied to enhance natural accuracy.

Method	Average		Worst	
	acc_{nat}	acc_{AA}	acc_{nat}	acc_{AA}
CFA	55.57	24.56	23.00	2.00
WAT	55.99	22.89	19.00	3.00
FAAL	55.87	23.57	21.00	3.00
Ours	55.88	23.14	19.00	2.00
Ours*	59.42	26.54	23.00	4.00

4 Experimental Results

4.1 Setup

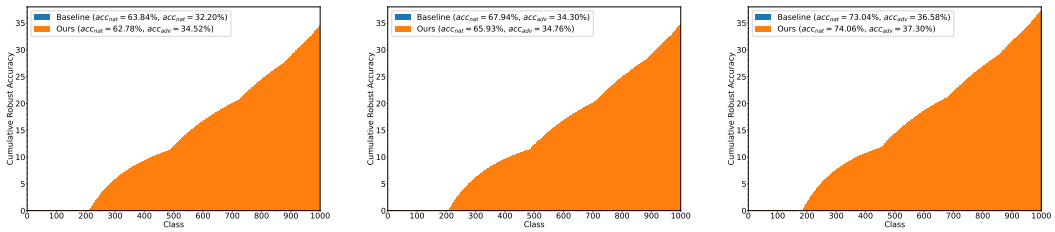
We assessed our proposed approach on the CIFAR-10, CIFAR-100, and ImageNet datasets under white-box attack scenarios with L_∞ norm constraints. Robustness was evaluated using the AutoAttack [20] benchmark, setting the perturbation bound to $\epsilon = 8/255$ for CIFAR-10 and CIFAR-100, and $\epsilon = 4/255$ for ImageNet. For a comprehensive comparison, baseline methods, CFA [17], WAT [16], FAAL [19], TRADES [9], and AWP [21], were reproduced using their official repositories with default settings. Our framework was implemented based on the official AWP repository. For experiments without AWP, we disabled the weight-perturbation mechanism, retaining the TRADES-style objective implemented in the same codebase. The same implementation was used for CIFAR-10 and CIFAR-100. For our post-training extension, we further applied our complete framework to pretrained checkpoints through a lightweight fine-tuning stage. In particular, AWP† denotes a pretrained AWP model fine-tuned with our complete framework for only five epochs, without repeating full adversarial training. The initial fine-tuning learning rate was set to the learning rate reached after the first scheduled decay in the original training configuration, followed by another decay at the third fine-tuning epoch.

4.2 Fairness on CIFAR-10 dataset

Table 1 reports the standard natural accuracy (acc_{nat}) and robust accuracy under AutoAttack (acc_{AA}) for the CIFAR-10 dataset. Our sample-level approach achieves a substantial improvement in worst-class robust accuracy over standard TRADES and WAT. While CFA reports a higher worst-class robust accuracy (28.10%), this comes at the cost of severe convergence instability, causing its average natural accuracy to plummet to 71.17%, markedly lower than competing methods. This demonstrates that our class-agnostic method can effectively enforce robust fairness without suffering catastrophic degradation in standard performance.

Prior studies highlight that adversarially trained models typically overfit immediately following the first learning rate decay [11]. To investigate the interplay between fairness and overfitting, we integrated Adversarial Weight Perturbation (AWP) [21], a layer-wise regularization technique, into our framework. Interestingly, applying AWP alone natively achieves a worst-class robustness of 25.30%, outperforming existing class-wise frameworks like WAT. This strongly validates our core hypothesis: introducing regularization to mitigate robust overfitting intrinsically alleviates fairness degradation. Integrating AWP with our proposed sample-level reweighting (Ours+AWP) further amplifies this synergy, boosting worst-class robustness to 28.60% while maintaining excellent average accuracy.

More importantly, as established in previous sections, robust overfitting remains the primary bottleneck; however, performance can be remarkably enhanced through a purely post-training approach. By applying only our Dirichlet probability redistribution to a pre-trained AWP model (AWP†), a lightweight fine-tuning process significantly elevated worst-class robustness to 29.00%, eliminating the requirement for costly training from scratch.



(a) ResNet-50 [22] (b) ViT [23] (c) Swin [23]

Figure 4: Performance Evaluation on ImageNet dataset using ResNet-50, Vit, and Swin models.

4.3 Fairness on CIFAR-100 dataset

While existing works primarily benchmark robust fairness on CIFAR-10, we extend our evaluation to CIFAR-100 to test scalability. Because both datasets share the same overall volume of training images, the data per class in CIFAR-100 is ten times sparser, making it a substantially more challenging benchmark for class-wise statistics.

Table 2 presents the standard and robust accuracies on CIFAR-100, as reported by FAAL. Row marked with an asterisk (*) indicates additional configurations applied to enhance natural accuracy, details of which will be discussed later. While FAAL successfully improves robust fairness on simpler datasets, its reweighting algorithm, based on Distributionally Robust Optimization (DRO), faces severe scalability challenges as the number of classes increases. As, each CIFAR-100 test class contains exactly 100 images, the improvement in the worst-class robust accuracy is only about 1%, a margin highly susceptible to evaluation noise and statistically insignificant.

A closer diagnosis of the CIFAR-100 results revealed a critical bottleneck: the class with the lowest robust accuracy also exhibited severely degraded standard accuracy.In this case, a purely post-training approach is not suitable, as the model's fundamental capacity to recognize natural features for these classes is already compromised. A more holistic strategy must explicitly boost the standard accuracy of these challenging samples first. However, our proposed framework provides the inherent flexibility to adaptively adjust optimization importance. By implementing an enhanced hybrid strategy (denoted as Ours*), which deliberately reduces the effective weight of the adversarial loss term $\beta\omega(x, x', y)$ to focus more heavily on natural examples, we prevent smaller-capacity models like ResNet-18 from being overwhelmed by rigid adversarial constraints. As can be seen, this targeted flexibility significantly boosts worst-class robust accuracy to 4.00% and average natural accuracy to 59.42%, effectively counteracting the data volume bottleneck.

4.4 Fairness on ImageNet dataset

The ImageNet dataset contains 1,000 classes, each with only a limited number of evaluation images (typically 5 to 10). As a result, class-wise accuracy becomes an unreliable metric, since each class can only take on a small number of discrete values. Moreover, for most pre-trained models, approximately 200–300 classes have no correctly classified samples under adversarial attacks. Given these challenges, the problem formulation needs to be adjusted. Alternative, the performance is accessed by the following metrics:

1. Cumulative accuracy from the worst-performing classes, defined as the total number of correctly classified adversarial examples from the worst class up to the i-th class, divided by the total number of evaluated examples over the same range.
2. The number of classes with zero robust accuracy,

Figure 4 visualizes the performance of our post-training approach applied to ResNet-50 [22], ViT [23], and Swin [23] architectures, utilizing pre-trained weights downloaded from the public RobustBench leaderboard [24]. As shown, the baseline models suffer from approximately 220 to 250 classes with completely zero robust accuracy. Remarkably, our lightweight post-training method systematically resuscitates these highly vulnerable classes, reducing the number of zero-robustness classes down to approximately 180 to 200. Furthermore, our proposed approach achieves a significantly higher

cumulative robust accuracy curve, dominating particularly among the worst-performing classes, while simultaneously yielding a slight increase in overall average robust accuracy.

4.5 Comparison to Existing Frameworks

The progressive empirical results highlight three fundamental advantages of our sample-level framework:

1. Zero Validation Data Cost: To compute historical class statistics, existing methods demand a dedicated validation set that scales linearly with the number of classes. Reserving just 100 samples per class sacrifices 20% of CIFAR-10 and 8% of ImageNet. Our sample-level approach is entirely self-contained, preserving 100% of the dataset for actual optimization.

2. Immunity to Hardware and Batch-Size Bottlenecks: Methods like FAAL attempt to bypass validation splits by calculating class weights within the mini-batch. However, on standard GPUs, memory constraints limit batch sizes to 200–400 samples. When applied to 1,000-class datasets, a single mini-batch covers only a fraction of the label space, rendering margin estimations mathematically unstable. Our framework evaluates risks individually per sample, remaining completely agnostic to batch size constraints.

3. Fine-Grained Resolution and Flexibility: Class-wise reweighting inherently fails to differentiate between a vulnerable boundary image and an unlearnable outlier within the same class. Our self-comparison mechanism provides fine-grained, instance-level adjustments. Depending on the dataset's complexity and data volume, our framework offers the flexibility to explicitly penalize overconfidence via post-training (for CIFAR-10 and ImageNet) or dynamically shift optimization focus toward natural learnability (for CIFAR-100), establishing a truly scalable and versatile paradigm for robust fairness.

5 Conclusion

In this paper, we demonstrated that class-wise disparities in adversarial training are fundamentally driven by robust overfitting. Because standard methods waste capacity on unlearnable outliers and already-safe data, we proposed a novel, sample-level tri-regime optimization framework that dynamically assigns optimization priorities by evaluating natural and boundary risks entirely at the instance level. By completely circumventing class-wise tracking and validation splits, our approach fundamentally solves the scalability bottlenecks that hinder existing methods. Our framework also provides the inherent flexibility to adjust optimization importance based on dataset properties, whether by dynamically shifting focus toward natural learnability or explicitly penalizing overconfidence.

Extensive evaluations validate that our approach achieves state-of-the-art worst-class robustness on CIFAR-10 and CIFAR-100. Most notably, we demonstrate that a purely post-training application of our sample-level regularization is highly effective. This lightweight strategy successfully resuscitated dozens of "zero-robustness" classes on ImageNet, establishing our framework as the first fair adversarial training method to successfully scale to this massive dataset.

Despite these advances, scaling to complex datasets exposes a serious underlying issue: rigidly enforcing uniform adversarial constraints on inherently difficult samples inevitably degrades their fundamental learnability. Because this severe degradation compromises the practical utility of the model, the original definition of fair adversarial training must be revised. Moving forward, the objective must evolve beyond pursuing strict class-wise parity. Instead, it should adaptively balance natural learnability with instance-level robustness, paving the way for a more holistic, scalable, and practical AI security paradigm.

Acknowledgments and Disclosure of Funding

We would like to express our gratitude to the National Science and Technology Council (NSTC) for financially supporting this research under grant numbers 115-2221-E-007-110-MY3 and 114-2221-E-007-076-MY3.

We also thank the National Center for High-performance Computing (NCHC) of the National Institutes of Applied Research (NIAR), Taiwan, for providing advanced research infrastructure and invaluable support.

References

[1] Xiaoyan Jiang, Shuihua Wang, and Yudong Zhang. Vision transformer promotes cancer diagnosis: A comprehensive review. *Expert Systems with Applications*, 252:124113, 2024.

[2] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Debarag Banerjee, Pooja Singh, Arjun Avadhanam, and Saksham Srivastava. Benchmarking llm powered chatbots: methods and metrics. *arXiv preprint arXiv:2308.04624*, 2023.

[4] Sergio Paniego, Enrique Shinohara, and Josemaria Canas. Autonomous driving in traffic with end-to-end vision-based deep learning. *Neurocomputing*, 594:127874, 2024.

[5] Lijun Chi, Mounira Msahli, Qingjie Zhang, Han Qiu, Tianwei Zhang, Gerard Memmi, and Meikang Qiu. Adversarial attacks on autonomous driving systems in the physical world: a survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] Taifeng Liu, Chao Yang, Xinjing Liu, Ruidong Han, and Jianfeng Ma. Rpau: Fooling the eyes of uavs via physical adversarial patches. *IEEE Transactions on Intelligent Transportation Systems*, 25(3):2586–2598, 2023.

[7] Kien Nguyen, Tharindu Fernando, Clinton Fookes, and Sridha Sridharan. Physical adversarial attacks for surveillance: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[9] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[10] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International conference on machine learning*, pages 36246–36263. PMLR, 2023.

[11] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pages 8093–8104. PMLR, 2020.

[12] Erh-Chung Chen and Che-Rung Lee. Data filtering for efficient adversarial training. *Pattern Recognition*, 151:110394, 2024.

[13] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021.

[14] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

[15] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR, 2021.

[16] Boqi Li and Weiwei Liu. Wat: improve the worst-class robustness in adversarial training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14982–14990, 2023.

[17] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8193–8201, 2023.

[18] Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fariness via balance adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15161–15169, 2023.

[19] Yanghao Zhang, Tianle Zhang, Ronghui Mu, Xiaowei Huang, and Wenjie Ruan. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24746–24755, 2024.

[20] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[21] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

[22] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[23] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 35:18599–18611, 2022.

[24] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our main contributions in the abstract and introduction sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The method we propose needs to be adjusted as necessary to better adapt to various datasets.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We explain the design philosophy in the main content.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setups are presented accordingly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released once it is accepted.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setups are presented accordingly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Certain results are sourced from public works or benchmarks that originally lacked error bars. However, we have addressed this limitation specifically within the CIFAR-100 dataset experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We address the computational budget constraints regarding the ImageNet dataset.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This work does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLM is used only for writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.