# ENHANCING ROBUSTNESS THROUGH REPAIR-ORIENTED HONEYPOTS AND SELECTIVE KNOWLEDGE AGGREGATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A model with established adversarial robustness can remain vulnerable to an encountered attack. Targeted improvement requires useful repair data and a learning procedure that limits damage to existing capabilities. We introduce HoneyTeacher, a general framework that couples repair-oriented honeypot induction with selective knowledge aggregation. The production model's estimated repair benefit guides honeypot configuration, shaping attack interactions to yield training material with greater downstream learning value. A material teacher learns from these data; aggregation then selects supervision from this teacher and the original model according to their joint correctness, consolidating complementary corrections while retaining existing strengths. We evaluate score-based black-box and gradient-based white-box instantiations on all 10,000 CIFAR-10 test images. Under Square and PGD attacks, respectively, target robust accuracy improves by 12.4% and 8.7% relative to the original model, and by 17.1% and 13.5% relative to ordinary-material direct learning. On the secondary Square evaluation, the white-box model improves accuracy by 24.0% relative to the original model and 34.2% relative to ordinary-material direct learning. Clean accuracy changes by +0.17 and -1.26 percentage points relative to the original model in the black-box and white-box settings, respectively, with all monitored capabilities meeting the adopted retention criteria. Induced material improves robustness under matched direct learning, and aggregation adds further gains; together, they reverse the net target-robustness losses observed with ordinary direct learning in both settings.

## 1 INTRODUCTION

Adversarial training can make a model resistant to input perturbations designed to cause prediction errors within a specified threat model (Madry et al., 2018). Nevertheless, a model with established robust capabilities may remain insufficiently resistant to a particular attack encountered during deployment (Dai et al., 2025). The task is to strengthen the production model against that attack while retaining the capabilities it already possesses. An effective update should achieve substantial targeted improvement with little degradation of existing robustness and clean predictive performance.

The attacker's search offers a source of inputs from which the production model can learn. Because this search depends in part on the exposed model and its feedback (Chen et al., 2022), the defender has an opportunity to influence the material it acquires. We consider a setting where an upstream mechanism identifies adversarial activity and routes it to a honeypot: a defender-configured model that interacts with the attacker, while the production model remains the object of improvement. Configuring this interaction around the production model's learning needs could yield more useful material than ordinary exposure. This leads to our research question: *How can we shape attack interactions to acquire more useful learning material and translate its information into greater robustness against the encountered attack with little damage to existing capabilities?*

Acquiring such material requires connecting attack search to downstream learning value. An input's effectiveness against the exposed model does not directly quantify how much learning from it would strengthen the production model. Indeed, emphasizing or more closely fitting difficult adversarial examples need not improve robust generalization (Kanai et al., 2023; Lee et al., 2025). The challenge



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

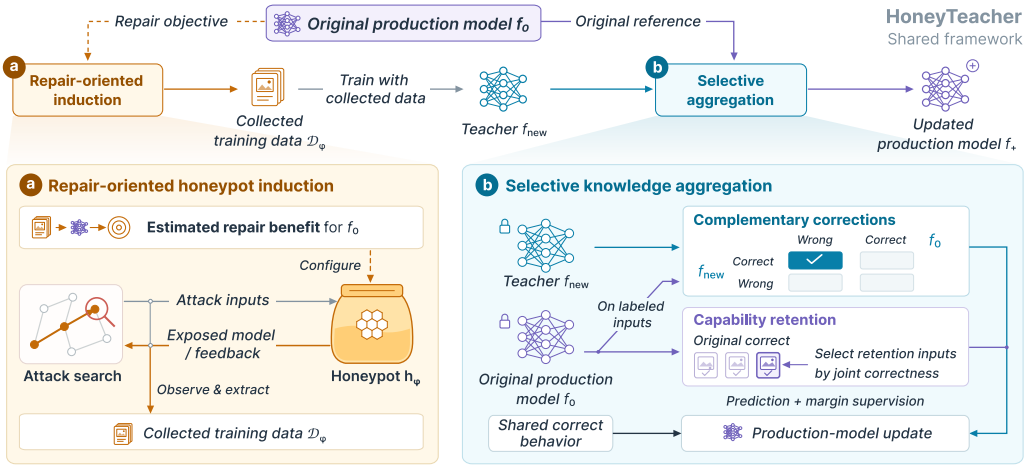


Figure 1: **HoneyTeacher framework.** **(a)** Repair-oriented induction uses the estimated learning benefit for the original production model $f_0$ to guide honeypot interactions toward training data with greater repair value. These data train a material teacher $f_{\text{new}}$. **(b)** The shared aggregation core selects teacher supervision by joint correctness to consolidate complementary corrections and retain existing correct behavior. The check mark highlights inputs misclassified by $f_0$ but correctly classified by $f_{\text{new}}$. Together, the two stages target robustness gains with limited capability degradation. Score-based and gradient-based access instantiate the same framework.

is to translate the production model's repair needs into a honeypot design that elicits material with greater learning value: candidates must be assessed by their anticipated contribution to targeted improvement, and this assessment must shape the feedback available during attack search.

Realizing that learning value also requires selective absorption of the acquired information. Adaptation to new attacks can erode existing robustness (Zhou & Hua, 2024; Dai et al., 2025), and even an update with higher overall robust accuracy can introduce failures on previously protected inputs (Angioni et al., 2025; Yan et al., 2021). Simply reducing the update magnitude constrains both beneficial and harmful changes. Conversely, indiscriminately preserving the original model's predictions can reinforce the weaknesses that adaptation is intended to address. The challenge is therefore to identify which information supplies complementary corrections and where existing behavior should be retained, so that useful material translates into targeted gains with limited damage.

We introduce HoneyTeacher, a general framework connecting repair-oriented honeypot induction with selective knowledge aggregation (Figure 1). The production model's repair needs guide both material acquisition and knowledge aggregation. Repair-oriented utility shapes honeypot feedback to steer attack search toward material with greater anticipated learning value than ordinary exposure.

HoneyTeacher then trains a material teacher on the acquired material and selectively combines its supervision with that of the original model. On training inputs, it absorbs the material teacher's corrections to the original model's failures while prioritizing retention where previously correct behavior is at risk of degradation. This distinguishes which corrections to absorb from which behaviors to preserve, addressing the tension left unresolved by uniform update shrinkage. We evaluate the shared framework through score-based black-box and gradient-based white-box instantiations.

We evaluate HoneyTeacher on all 10,000 CIFAR-10 test images. Under black-box Square and white-box PGD evaluation, the complete framework improves target robust accuracy by 12.4% and 8.7% relative to the original production model, and by 17.1% and 13.5% over ordinary-material direct learning, respectively. The white-box model also achieves relative gains of 24.0% over the original model and 34.2% over ordinary-material direct learning on the secondary Square evaluation. All monitored capabilities meet the adopted retention criteria. Relative to the original production model, clean accuracy changes by +0.17 and -1.26 percentage points in the black-box and white-box settings, respectively. In both settings, induced material yields higher target robustness under matched

direct learning, and selective aggregation provides further gains. Together, these stages turn the net target-robustness losses observed with ordinary learning into net gains.

## 2 Related Work

### 2.1 Honeypot-Based Defenses and Attack Interaction

Model-interface defenses use attack interactions for detection and feedback manipulation. Trapdoor signatures and similarities across successive queries support detection in honeypot defenses and Blacklight, respectively (Shan et al., 2020; Li et al., 2022). Feedback manipulation changes the information exposed to attackers (Wang et al., 2025): AAA perturbs scores to mislead evasion search, whereas Prediction Poisoning perturbs predictions to impair learning in model stealing (Chen et al., 2022; Orekondy et al., 2020).

Honeypots also support recovery and training: NeuroPots attracts bit flips to honey neurons for parameter restoration (Liu et al., 2023), while honeypot modules capture backdoor correlations to downweight suspected poisoned examples during language-model fine-tuning (Tang et al., 2023). These uses address attack detection, disruption, or harmful effects, but can leave the defender without material tailored to the production model's repair needs. HoneyTeacher uses repair utility to guide honeypot induction toward material with greater learning value than ordinary exposure.

### 2.2 Adversarial Material and Robust Learning

Work on adversarial material spans example construction, weighting, and augmentation. Friendly adversarial training uses early-stopped attacks to construct examples more amenable to learning (Zhang et al., 2020). Reweighting uses geometric vulnerability or distributionally robust optimization (Zhang et al., 2021; Sow et al., 2024), while generative approaches expand training distributions with synthetic data (Gowal et al., 2021; Sehwag et al., 2022; Wang et al., 2023).

Attack difficulty alone does not capture learning value. Excessive fitting to hard adversarial instances is linked to memorization and robust overfitting (Liu et al., 2024; Lee et al., 2025); selecting difficult examples does not consistently outperform random selection in subset adversarial training (Losch et al., 2024). Learning-aware selection considers contributions to an optimization objective (Killamsetty et al., 2021). Meta-reweighting uses gradient effects on validation loss for noisy or imbalanced data (Ren et al., 2018), while adversarial coreset selection approximates the full robust-training gradient with weighted subsets (Dolatabadi et al., 2023).

Success against the exposed model need not imply repair value. HoneyTeacher uses the estimated reduction in the production model's repair loss from learning on candidate material to guide honeypot feedback, aiming to elicit material with greater targeted learning value than ordinary exposure.

### 2.3 Knowledge Aggregation and Capability Retention

Adversarial distillation transfers robustness to students through teacher supervision (Goldblum et al., 2020; Zi et al., 2021). IAD adapts teacher trust to the input, while balanced multi-teacher adversarial distillation coordinates clean and robust teachers (Zhu et al., 2022; Zhao et al., 2024). To mitigate forgetting when adapting to new attacks, continual defenses use pseudo-replay, self-distillation, and logit-space regularization (Zhou & Hua, 2024; Dai et al., 2025). Adaptive multi-teacher self-distillation also preserves knowledge across adversarial-training stages (Wang & Ding, 2025).

Higher average performance can conceal new failures on previously correct inputs. Positive-congruent training and weight interpolation reduce these negative flips in image and text classification, respectively (Yan et al., 2021; Schumann et al., 2024). Robustness-congruent adversarial training (RCAT) combines distillation from an improved source model with preservation on inputs correctly classified by the old model (Angioni et al., 2025). For class expansion, margin preserving training (MPT) couples margin calibration with distillation from old and new models, weighting supervision by each teacher's correctness (Ricci et al., 2026).

These update methods address how to incorporate improved models with fewer regressions, but not how to shape external attack interactions to acquire knowledge tailored to the production model's

repair needs. HoneyTeacher couples repair-oriented honeypot induction with selective knowledge aggregation: a teacher trained on induced material supplies complementary corrections, which are consolidated while retaining the production model's existing correct behavior.

## 3 Problem Setup

We first specify HoneyTeacher's interaction setting: attacker access, defender control, and observations available for learning. We then define material value relative to ordinary exposure under a common learner, alongside target robustness gain and capability loss of the updated model.

### 3.1 Interaction Setting and Information Access

We consider a production classifier $f_0 : \mathcal{X} \to \mathbb{R}^K$ with established robust capabilities but insufficient resistance to an encountered attack. Here $\mathcal{X}$ is the input space, and $f_0(x)$ gives logits for $K$ classes. We take upstream detection and routing as given: identified adversarial interactions are routed to a defender-configured honeypot $h_\phi$, whose configuration $\phi$ controls the exposed model or feedback. The production model $f_0$ remains the object of repair.

For a source input $x$, an attacker $\mathcal{A}$ searches within an admissible set $\mathcal{N}(x)$ under a specified interaction or computation budget. We evaluate the general framework through score-based black-box access, where the attacker queries model scores, and gradient-based white-box access, where it accesses the exposed model and differentiates its forward computation. In the gradient-based instantiation, the defender observes the inputs queried during acquisition.

The defender extracts learning material $\mathcal{D}_\phi$ from observations $\mathcal{O}_\phi$ available under the acquisition protocol. Repair also uses designated training and reference data.

We assess induction by comparing $h_\phi$ with ordinary exposure $h_{\mathrm{ord}}$, which need not coincide with $f_0$. Within each evaluated setting, both arms share source inputs and attack-budget limits, with stopping rules specified by the acquisition protocol. The updated production classifier $f_+$ is evaluated using attacks newly generated against it on inputs held apart from material acquisition and repair training.

### 3.2 Learning Objective

Within each access setting, a common direct learner $\mathcal{U}$ produces $f_{\mathrm{ord}} = \mathcal{U}(f_0, \mathcal{D}_{\mathrm{ord}})$ and $f_{\mathrm{new}} = \mathcal{U}(f_0, \mathcal{D}_\phi)$ from ordinary and induced material, with training and reference data, initialization, and optimization settings held fixed. Selective aggregation uses the material teacher $f_{\mathrm{new}}$ to obtain $f_+$.

An evaluation condition $c$ specifies a labeled-input distribution $P_c$, an admissible perturbation set, and a fixed-budget attack protocol. Let $s_c(f; x, y)$ indicate correct classification of the clean input and resistance to the specified attacks generated against $f$ itself. For the target condition $t$, representing the encountered attack and input population to be strengthened, we define robust accuracy and target gain as

$$R_c(f) = \mathbb{E}_{(x,y)\sim P_c}[s_c(f; x, y)], \qquad G_t(f_+) = R_t(f_+) - R_t(f_0).$$

The expectation includes attack randomness where applicable.

The incremental learning value of induced material under the common learner is

$$V_{\mathcal{U}}(\mathcal{D}_\phi; \mathcal{D}_{\mathrm{ord}}) = R_t(f_{\mathrm{new}}) - R_t(f_{\mathrm{ord}}).$$

The local utility used to configure the honeypot guides material construction; $V_{\mathcal{U}}$ measures its downstream learning value. Comparing $f_+$ with $f_{\mathrm{new}}$ measures the aggregation-stage change, whereas comparing $f_+$ with $f_{\mathrm{ord}}$ evaluates the complete framework against ordinary-material learning. All comparisons use the same evaluation condition and population.

For each designated retention condition $c$ (target or additional capabilities), paired damage is

$$H_c(f_+; f_0) = \Pr[s_c(f_0; x, y) = 1,\ s_c(f_+; x, y) = 0]\,.$$

This probability uses $P_c$ and the specified paired attack protocol, including attack randomness, with the full evaluation population as its denominator. Under this condition and protocol, the net change in $R_c$ equals the repair fraction (previously unsuccessful cases becoming successful) minus $H_c$. Our empirical retention criterion uses net capability changes; paired repairs and damage are reported separately to characterize which behaviors are gained and lost.

# 4 HONEYTEACHER

HoneyTeacher couples repair-oriented honeypot induction with selective knowledge aggregation through a material teacher (Figure 1). Induction shapes the exposed model or feedback to acquire material $\mathcal{D}_\phi$ with greater anticipated repair value than ordinary exposure. A material teacher $f_{\text{new}}$ learns from $\mathcal{D}_\phi$; aggregation combines its complementary corrections with correct behavior retained from $f_0$ to produce $f_+$. Score-based and gradient-based instantiations adapt acquisition and material learning to their interfaces while sharing the aggregation objective in Section 4.3.

## 4.1 REPAIR-ORIENTED HONEYPOT INDUCTION

Repair-oriented induction configures the exposure $h_\phi$ using downstream learning benefit, aiming to acquire material with greater repair value for $f_0$ than ordinary exposure.

In the score-based instantiation, a class-specific reference direction estimates each candidate's repair utility. For a source $x$, its unique predicted label $k = \arg\max_j f_0(x)_j$ remains fixed throughout acquisition and material learning. Let $\theta_0$ be the parameters of $f_0$, and let $J_{\text{ref},k}(\theta)$ be the mean cross-entropy on adversarial reference inputs with ground-truth class $k$, generated against $f_0$ and held fixed during differentiation. For a candidate $z$ from this source, define

$$g_{\text{ref},k} = \nabla_\theta J_{\text{ref},k}(\theta_0), \qquad g_{z,k} = \nabla_\theta \ell(f_\theta(z), k)|_{\theta=\theta_0},$$

where $\ell$ is cross-entropy. The local utility is

$$u(z;k) = \langle g_{\text{ref},k}, g_{z,k} \rangle.$$

For the hypothetical learning step $\theta_0 - \eta g_{z,k}$, with $\eta > 0$, an $L_k$-Lipschitz gradient of $J_{\text{ref},k}$ along the update segment gives

$$J_{\text{ref},k}(\theta_0) - J_{\text{ref},k}(\theta_0 - \eta g_{z,k}) = \eta u(z;k) + r_{z,k},$$
$$|r_{z,k}| \le \frac{L_k \eta^2}{2} \|g_{z,k}\|^2.$$

Thus, $\eta u(z;k)$ predicts the first-order reduction in the assigned class's repair loss. Encoding this utility in the exposed scores makes estimated learning benefit part of the attacker's search objective.

In the gradient-based instantiation, repair-dependent feedback enters the exposed forward computation and shapes the gradients available to attack search. Feedback parameters are selected through acquisition and learning: each configuration produces material, a fixed learning procedure trains its material teacher $f_{\text{new}}$, and the teacher's robustness on a designated design set guides the search.

In both settings, realized material value is assessed by the matched direct-learning comparison in Section 3.2. Section 4.4 specifies the feedback constructions; Section 4.2 describes how acquired material trains the teacher for selective knowledge aggregation.

## 4.2 FROM INTERACTION MATERIAL TO LEARNING INFORMATION

Observations $\mathcal{O}_\phi$ yield material $\mathcal{D}_\phi$, whose candidates inherit labels $\tilde{y}_i$ and confidence weights $w_i$ fixed before acquisition at the source inputs initiating attack sessions. Within each access setting, ordinary and induced material share extraction rules and allocation; matched learning from $f_0$ produces $f_{\text{ord}}$ and the material teacher $f_{\text{new}}$ (Section 3.2). For the scheduled material positions $\mathcal{B}_\mathcal{D}$ at learning step $t$, the confidence-weighted material loss is

$$\mathcal{L}^{(t)}_{\text{mat}}(\theta) = \frac{1}{|\mathcal{B}_\mathcal{D}|} \sum_{i \in \mathcal{B}_\mathcal{D}} w_i \, \ell(f_\theta(z_i^{(t)}), \tilde{y}_i),$$

where $\ell$ is cross-entropy and $z_i^{(t)}$ is the retained input or its prescribed training-time transformation. The denominator counts scheduled positions, including repetitions, rather than summing confidence weights. Both learners also distill $f_0$ on fixed protection views it classifies correctly. Appendix C specifies material extraction and the complete training procedures and objectives.

After learning, we freeze $f_{\text{new}}$ and retain the unchanged $f_0$ as the reference for existing capabilities. Their joint correctness on labeled aggregation inputs selects supervision for aggregation (Section 4.3). The ordinary-material baseline $f_{\text{ord}}$ remains unaggregated.

### 4.3 Selective Knowledge Aggregation with Capability Retention

Selective knowledge aggregation consolidates corrections acquired through induction while retaining complementary capabilities of $f_0$. Both $f_0$ and $f_{\text{new}}$ remain frozen, and the student $f_\theta$ is initialized from $f_{\text{new}}$ in both access settings.

For a labeled input $(z, y)$, define the logit margin $m_f(z, y) = f(z)_y - \max_{j \neq y} f(z)_j$ and correctness indicator $c_f(z, y) = \mathbf{1}\{m_f(z, y) > 0\}$. Let $c_0 = c_{f_0}(z, y)$ and $c_{\text{new}} = c_{f_{\text{new}}}(z, y)$. Inputs correctly classified by at least one teacher form three disjoint groups:

$$\begin{aligned} M_{\text{old}} &= c_0(1 - c_{\text{new}}), & t_{\text{old}} &= f_0, \\ M_{\text{new}} &= (1 - c_0)c_{\text{new}}, & t_{\text{new}} &= f_{\text{new}}, \\ M_{\text{shared}} &= c_0 c_{\text{new}}, & t_{\text{shared}} &= f_{\text{new}}. \end{aligned}$$

Here $t_g$ is the supervising teacher for group $g$. Old-only supervision recovers original behaviors missing from the material teacher; new-only supervision consolidates its complementary corrections; shared-correct supervision reinforces jointly correct decisions.

For each eligible input, prediction distillation is combined with a penalty for falling below the selected teacher's capped margin:

$$\begin{aligned} \ell_g(z, y) = T^2 \,\mathrm{KL}\Big(p^T_{t_g}(z) \,\|\, p^T_{f_\theta}(z)\Big) \\ + \big[\min\{m_{t_g}(z, y), \tau\} - m_{f_\theta}(z, y)\big]_+ , \end{aligned}$$

where $p^T_f(z) = \mathrm{softmax}(f(z)/T)$, $T > 0$ is the distillation temperature, $[a]_+ = \max\{a, 0\}$, and $\tau > 0$ caps the reference margin. Teacher predictions and group assignments are detached during student optimization.

Each aggregation anchor supplies a clean view, fixed adversarial views generated against the teachers, and an online adversarial view generated against the student. Let $\mathcal{V}$ index the views, $\mathcal{B}_v$ denote the corresponding minibatch, and $\mathcal{G} = \{\text{old}, \text{new}, \text{shared}\}$. The shared aggregation objective is

$$\mathcal{L}_{\text{sel}} = \sum_{g \in \mathcal{G}} \frac{1}{K|\mathcal{V}|} \sum_{v \in \mathcal{V}} \sum_{k=1}^{K} \mathrm{Mean}_{\substack{(z,y) \in \mathcal{B}_v \\ y=k,\ M_g(z,y)=1}} \ell_g(z, y).$$

Group losses are averaged over eligible inputs within each view–class cell and then over all $K|\mathcal{V}|$ cells, with empty cells contributing zero. Cell weights therefore do not depend on eligible counts.

The trained student becomes $f_+$. View construction and the white-box extension using correct clean-teacher targets are specified in Appendices D.2 and D.4.

### 4.4 Shared Structure and Instantiations

Both instantiations use the same production model $f_0$ and all $K$ classes. Ordinary exposure uses $f_0$ unchanged.

**Score-based access.** Sources are balanced across production-predicted classes, inheriting fixed labels and confidence weights from $f_0$. Banded feedback encodes source-label cross-entropy as $-\log p_{\phi,k}(z) = t_k(z)$, with $0 < t_k(z) < \log 2$. The loss increases across production-boundary difficulty bands and with repair utility $u(z; k)$ within each band. Since $p_{\phi,k} > 1/2$, feedback preserves the source label after crossing the production boundary, concealing success and sustaining search. Ordinary Square sessions stop at first success; induced sessions use the common query ceiling, with realized costs reported separately. Both arms share utility screening and diversity selection.

**Gradient-based access.** One gradient step from $f_0$ on fixed adversarial reference inputs it misclassifies yields a frozen repair-reference model $f_{\text{r}}$. The loss reduction $U(z, k) = [\ell(f_0(z), k) - \ell(f_{\text{r}}(z), k)]/\eta_{\text{r}}$, scaled by the reference step size $\eta_{\text{r}}$, supplies the repair signal. Evaluated at the query's production-predicted class and normalized, it controls an increasing transformation of production logit gaps:

$$h_\phi(z)_j = -G_z(d_j(z)), \qquad d_j(z) = \max_k f_0(z)_k - f_0(z)_j, \qquad G'_z(d) > 0.$$

Class order is preserved while attack gradients differentiate the full repair-dependent computation. Fixed-budget acquisition–learning search on design data under retention constraints, followed by separate calibration, fixes the feedback before final acquisition. Appendices B and C detail the feedback constructions and the labeling, extraction, and refresh rules within each setting.

**Material learning and aggregation.** Within each setting, ordinary and induced learners share initialization, auxiliary data, processing, and optimization (Appendix C). Both use protection distillation. The induced teacher enters the shared aggregation mechanism in Section 4.3 to produce $f_+$; $f_{\mathrm{ord}}$ remains unaggregated.

# 5 Experiments

## 5.1 Experimental Setup

We evaluate HoneyTeacher's score-based black-box and gradient-based white-box instantiations from the same CIFAR-style ResNet-18 checkpoint (He et al., 2016) on all 10,000 CIFAR-10 test images (Krizhevsky, 2009) (1,000 per class). Each setting compares the original $f_0$, ordinary- and induced-material direct learners $f_{\mathrm{ord}}$ and $f_{\mathrm{new}}$, and aggregated $f_+$. Comparisons assess complete-framework gains ($f_+$ versus $f_0$ and $f_{\mathrm{ord}}$), material value under matched direct learning ($f_{\mathrm{new}}$ versus $f_{\mathrm{ord}}$), and aggregation effects ($f_+$ versus $f_{\mathrm{new}}$).

All attacks use an $\ell_\infty$ radius of $8/255$. The black-box target is Square with a 256-query search cap per restart and four restarts (Andriushchenko et al., 2020); the white-box target combines cross-entropy and difference-of-logits-ratio PGD, each with 100 steps and five restarts (Madry et al., 2018; Croce & Hein, 2020). We additionally evaluate clean accuracy in both settings, PGD-CE (50 steps, five restarts) and SimBA-DCT (1,024 queries, two restarts) (Guo et al., 2019) for black-box access, and FGSM (Goodfellow et al., 2015) and Square (1,024 queries, four restarts) for white-box access. Query caps apply per restart; budgets are matched across models within each setting.

Robust correctness requires correct predictions on clean inputs and all retained model-specific attack views. Repairs and damage count gains and losses of robust correctness relative to $f_0$; their difference is net repair. All accuracy and transition rates use the 10,000-image denominator. Class-stratified paired image bootstrap gives 95% confidence intervals conditional on the fixed models, material, and attacks. Appendix A details evaluation procedures, data roles, and study history.

## 5.2 End-to-End Framework Performance

HoneyTeacher improves the original production model's target robust accuracy by 12.4% under black-box Square and 8.7% under white-box PGD (Table 1). These correspond to gains of 5.72 percentage points (95% CI [5.16, 6.27]) and 2.50 points [2.01, 2.98], respectively. Against ordinary-material direct learning, relative improvements reach 17.1% and 13.5%, with absolute gains of 7.60 points [6.94, 8.25] and 3.72 points [3.21, 4.23].

The complete framework turns ordinary learning's net target-robustness losses into gains over $f_0$ (Figure 2). Net repair rises from $-188$ to $+572$ in the black-box setting and from $-122$ to $+250$ in the white-box setting, yielding 760 and 372 additional net correct outcomes over ordinary learning. Black-box HoneyTeacher achieves 1.86 times as many repairs while reducing damage by 76.2%. White-box HoneyTeacher achieves 6.37 times as many repairs (446 versus 70), with damage increasing by four images (196 versus 192).

Testing each model on attack witnesses from $f_0$, $f_{\mathrm{ord}}$, and $f_+$ corroborates the advantage: gains over ordinary learning remain 7.62 points [6.95, 8.27] for black-box access and 5.11 points [4.59, 5.64] for white-box access. Thus, the advantage persists beyond evaluations using only attacks generated against each model itself. Appendix E.1 specifies the pools and paired outcomes.

## 5.3 Learning Value of Induced Material

Induced material produces larger target-robustness gains under matched direct learning. Before aggregation, $f_{\mathrm{new}}$ exceeds $f_{\mathrm{ord}}$ by 5.27 percentage points in the black-box setting (95% CI [4.66, 5.89]) and 2.11 points in the white-box setting [1.80, 2.44] (Table 1). Black-box acquisition uses

Table 1: Joint evaluation on the same 10,000 test images. Accuracies are percentages; target repairs, damage, and net repair are counts relative to $f_0$, using attacks generated separately against each model. Bold marks the best updated-model result in each column and setting.

| Setting | Model | Target ↑ | Clean ↑ | Repairs ↑ | Damage ↓ | Net ↑ |
|---|---|---|---|---|---|---|
| *Black-box Square* | Original $f_0$ | 46.23 | 90.32 | – | – | 0 |
| | Ordinary direct $f_{\text{ord}}$ | 44.35 | 89.92 | 380 | 568 | −188 |
| | Induced direct $f_{\text{new}}$ | 49.62 | 90.15 | 544 | 205 | +339 |
| | HoneyTeacher $f_+$ | **51.95** | **90.49** | **707** | **135** | **+572** |
| *White-box PGD (CE+DLR)* | Original $f_0$ | 28.82 | 90.32 | – | – | 0 |
| | Ordinary direct $f_{\text{ord}}$ | 27.60 | **90.59** | 70 | 192 | −122 |
| | Induced direct $f_{\text{new}}$ | 29.71 | 90.24 | 142 | **53** | +89 |
| | HoneyTeacher $f_+$ | **31.32** | 89.06 | **446** | 196 | **+250** |

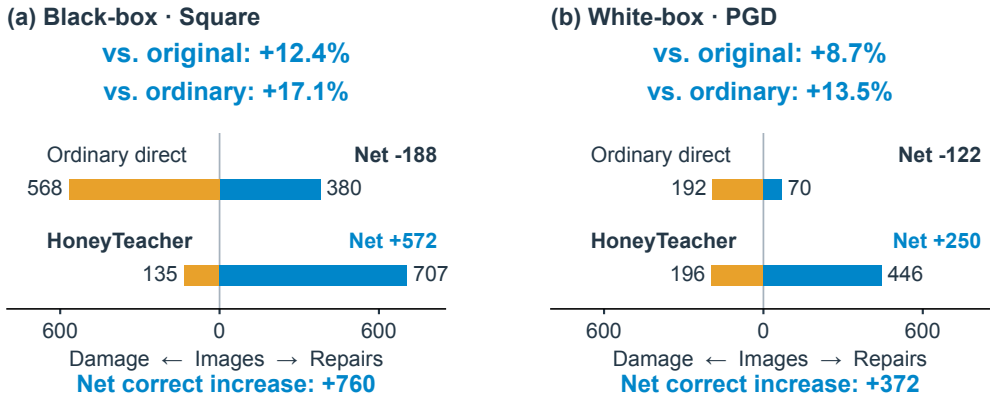


Figure 2: Complete-framework target gains and repair–damage balance on 10,000 images. Relative target-accuracy gains compare HoneyTeacher with each labeled reference. Bars count repairs (right) and damage (left) relative to $f_0$; net-correct increases compare HoneyTeacher with ordinary direct learning.

the same 256-query cap per source, with averages of 169 and 256 queries for ordinary and induced collection, respectively.

Relative to the observed full-framework accuracy advantage over $f_{\text{ord}}$, these pre-aggregation differences correspond to descriptive shares of 69.3% and 56.7%, respectively (Figure 3a,b). At the class level, induced-material direct learning achieves higher target accuracy than ordinary learning in all ten classes in both settings. The matched-learning comparison supports the central premise of repair-oriented induction: material elicited through the configured interaction yields higher target robustness than ordinary material before aggregation, providing a stronger material teacher for the subsequent stage.

### 5.4 Selective Aggregation and Capability Retention

Selective knowledge aggregation improves target robust accuracy beyond induced-material direct learning by 2.33 percentage points in the black-box setting (95% CI [1.94, 2.73]) and 1.61 points in the white-box setting [1.13, 2.08]. These increments correspond to descriptive shares of 30.7% and 43.3% of the observed full-framework advantage over ordinary learning, respectively (Figure 3a,b). Aggregation further improves target accuracy in nine of ten black-box classes and seven of ten white-box classes, with class-dependent gains and losses.

Figure 3c shows the final models' capability changes relative to $f_0$. Black-box clean, PGD, and SimBA accuracy change by $+0.17$, $+0.42$, and $+0.41$ points, with all three intervals including zero. White-box clean and FGSM changes are $-1.26$ points $[-1.64, -0.89]$ and $-0.77$ points

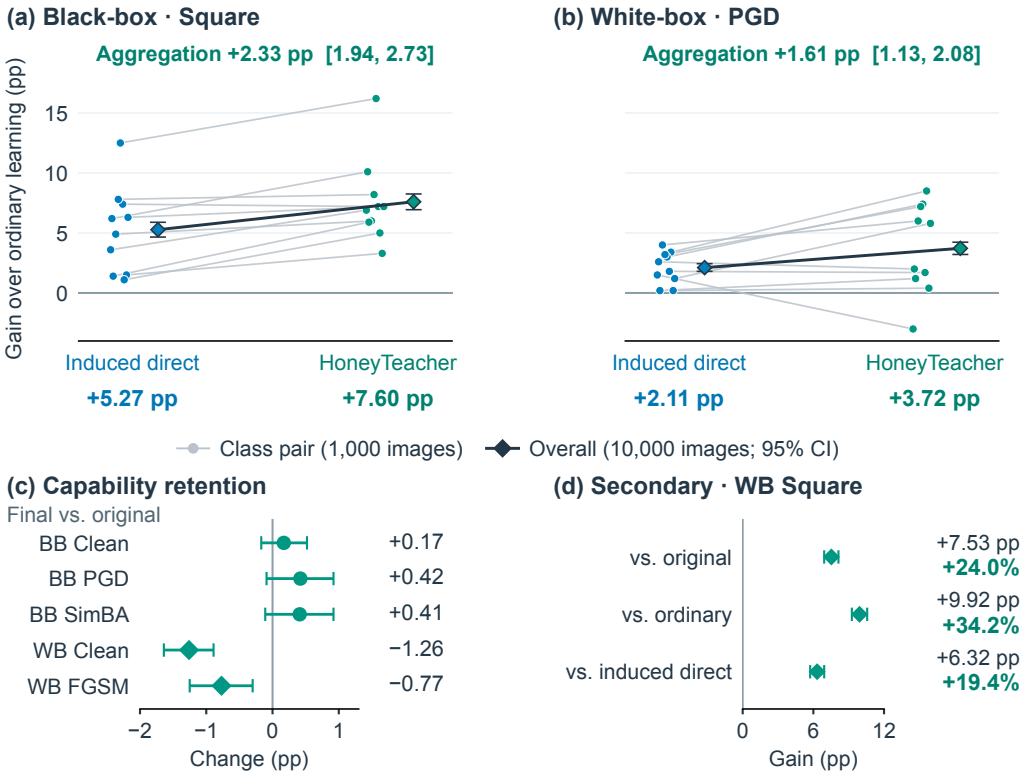


Figure 3: Stage gains, class-wise variation, and capability retention on 10,000 test images. (a,b) Gains over ordinary direct learning: small paired points represent ten classes of 1,000 images, and diamonds show overall effects. Header annotations report aggregation increments. (c) Final versus original accuracy changes. (d) Secondary white-box Square gains; relative percentages use each labeled reference. Whiskers and bracketed ranges give class-stratified paired-bootstrap 95% intervals conditional on fixed models, material, and attacks.

$[-1.25, -0.30]$, respectively. All monitored capabilities satisfy the adopted five-percentage-point net-degradation criterion against both $f_0$ and $f_{\text{ord}}$.

On the secondary Square evaluation (Figure 3d), the white-box model improves Square accuracy by 24.0% relative to $f_0$ and 34.2% relative to $f_{\text{ord}}$. The aggregation-stage gain over $f_{\text{new}}$ is 6.32 percentage points [5.73, 6.92], a 19.4% relative increase, complementing the primary PGD improvement. Appendix E.2 reports the complete capability contrasts; Appendix E.3 examines the white-box clean-target extension.

## 6 Conclusion

We presented HoneyTeacher, a general framework coupling repair-oriented honeypot induction with selective knowledge aggregation for targeted robustness enhancement. Production-model repair needs guide the acquisition of material with greater learning value than ordinary exposure provides, while aggregation integrates acquired corrections with complementary original-model capabilities. On the full CIFAR-10 test set, score-based black-box and gradient-based white-box instantiations show gains from both stages, turning ordinary learning's net robustness losses into gains while satisfying the adopted capability-retention criteria. These findings support jointly designing attack interactions and knowledge aggregation to improve robustness against encountered attacks while limiting measured degradation of existing capabilities.

### AI USE STATEMENT

We used generative AI tools to assist with language polishing of the manuscript. The authors take full responsibility for the final manuscript, including its technical content, claims, and reported results.

### REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the HoneyTeacher framework and learning objectives in Sections 3 and 4. Appendices B–D provide implementation details for repair-oriented feedback, material construction and teacher learning, and selective knowledge aggregation, including the associated losses and optimization settings. Appendix A specifies the production model, evaluation data, attack configurations, and statistical procedures. Appendix E provides detailed results, common-witness checks, and acquisition workload counts.

## REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision – ECCV 2020*, volume 12368 of *Lecture Notes in Computer Science*, pp. 484–501. Springer, 2020. doi: 10.1007/978-3-030-58592-1_29. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/4362_ECCV_2020_paper.php.

Daniele Angioni, Luca Demetrio, Maura Pintor, Luca Oneto, Davide Anguita, Battista Biggio, and Fabio Roli. Robustness-congruent adversarial training for secure machine learning model updates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7457–7469, 2025. doi: 10.1109/TPAMI.2025.3573237. URL https://doi.org/10.1109/TPAMI.2025.3573237.

Sizhe Chen, Zhehao Huang, Qinghua Tao, Yingwen Wu, Cihang Xie, and Xiaolin Huang. Adversarial attack on attackers: Post-process to mitigate black-box score-based query attacks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 14929–14943. Curran Associates, Inc., 2022. doi: 10.52202/068431-1086. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5fa29a2f163ce2020769eca8956e2d77-Abstract-Conference.html.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

Sihui Dai, Christian Cianfarani, Vikash Sehwag, Prateek Mittal, and Arjun Bhagoji. Adapting to evolving adversaries with regularized continual robust training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11954–12000. PMLR, 2025. URL https://proceedings.mlr.press/v267/dai25c.html.

Hadi M. Dolatabadi, Sarah M. Erfani, and Christopher Leckie. Adversarial coreset selection for efficient robust training. *International Journal of Computer Vision*, 131(12):3307–3331, 2023. doi: 10.1007/s11263-023-01860-4. URL https://link.springer.com/article/10.1007/s11263-023-01860-4.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3996–4003, 2020. doi: 10.1609/aaai.v34i04.5816. URL https://ojs.aaai.org/index.php/AAAI/article/view/5816.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://research.google/pubs/explaining-and-harnessing-adversarial-examples/.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4218–4233. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/21ca6d0cf2f25c4dbb35d8dc0b679c3f-Paper.pdf.

Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2484–2493. PMLR, 2019. URL https://proceedings.mlr.press/v97/guo19a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Sekitoshi Kanai, Shin'Ya Yamaguchi, Masanori Yamada, Hiroshi Takahashi, Kentaro Ohno, and Yasutoshi Ida. One-vs-the-rest loss to focus on important samples in adversarial training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15669–15695. PMLR, 2023. URL https://proceedings.mlr.press/v202/kanai23a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, 2021. doi: 10.1609/aaai.v35i9.16988. URL https://ojs.aaai.org/index.php/AAAI/article/view/16988.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Hyungyu Lee, Saehyung Lee, Ho Bae, and Sungroh Yoon. Regularizing hard examples improves adversarial robustness. *Journal of Machine Learning Research*, 26(20):1–48, 2025. URL https://www.jmlr.org/papers/v26/22-1428.html.

Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y. Zhao. Blacklight: Scalable defense for neural networks against query-based black-box attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2117–2134. USENIX Association, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/li-huiying.

Chen Liu, Zhichao Huang, Mathieu Salzmann, Tong Zhang, and Sabine Süsstrunk. On the impact of hard adversarial instances on overfitting in adversarial training. *Journal of Machine Learning Research*, 25(356):1–46, 2024. URL https://www.jmlr.org/papers/v25/22-0950.html.

Qi Liu, Jieming Yin, Wujie Wen, Chengmo Yang, and Shi Sha. NeuroPots: Realtime proactive defense against Bit-Flip attacks in neural networks. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 6347–6364. USENIX Association, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/liu-qi.

Max Losch, Mohamed Omran, David Stutz, Mario Fritz, and Bernt Schiele. On adversarial training without perturbing all examples. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d24189d251aa453b6d2eb3421ec48bcc-Abstract-Conference.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Prediction poisoning: Towards defenses against DNN model stealing attacks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyevYxHtDB.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Simone Ricci, Niccolò Biondi, Federico Pernici, and Alberto Del Bimbo. Mitigating negative flips via margin preserving training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):8721–8730, 2026. doi: 10.1609/aaai.v40i11.37825. URL https://ojs.aaai.org/index.php/AAAI/article/view/37825.

Raphael Schumann, Elman Mansimov, Yi-An Lai, Nikolaos Pappas, Xibin Gao, and Yi Zhang. Backward compatibility during data updates by weight interpolation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2846–2861. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.174. URL https://aclanthology.org/2024.eacl-long.174/.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WVX0NNVBBkV.

Shawn Shan, Emily Wenger, Bolun Wang, Bo Li, Haitao Zheng, and Ben Y. Zhao. Gotta catch'em all: Using honeypots to catch adversarial attacks on neural networks. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 67–83, 2020. doi: 10.1145/3372297.3417231. URL https://doi.org/10.1145/3372297.3417231.

Daouda Sow, Sen Lin, Zhangyang Wang, and Yingbin Liang. Doubly robust instance-reweighted adversarial training. In *International Conference on Learning Representations*, volume 2024, pp. 1274–1300, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/052d1262ec4174ec16356b7235cba1f7-Paper-Conference.pdf.

Ruixiang (Ryan) Tang, Jiayi Yuan, Yiming Li, Zirui Liu, Rui Chen, and Xia Hu. Setting the trap: Capturing and defeating backdoors in pretrained language models through honeypots. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73191–73210. Curran Associates, Inc., 2023. doi: 10.52202/075280-3199. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e7938ede51225b490bb69f7b361a9259-Abstract-Conference.html.

Xianglu Wang and Hu Ding. Exploring the forgetting in adversarial training: A novel method for enhancing robustness. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/572cd21bd5dea96b065476b77d21b3c6-Abstract-Conference.html.

Yixu Wang, Tianle Gu, Yan Teng, Yingchun Wang, and Xingjun Ma. HoneypotNet: Backdoor attacks against model extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8087–8095, 2025. doi: 10.1609/aaai.v39i8.32872. URL https://ojs.aaai.org/index.php/AAAI/article/view/32872.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36246–36263. PMLR, 2023. URL https://proceedings.mlr.press/v202/wang23ad.html.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14299–14308, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/

html/Yan_Positive-Congruent_Training_Towards_Regression-Free_Model_Updates_CVPR_2021_paper.html.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11278–11287. PMLR, 2020. URL https://proceedings.mlr.press/v119/zhang20z.html.

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. URL https://openreview.net/forum?id=iAX0l6Cz8ub.

Shiji Zhao, Xizhe Wang, and Xingxing Wei. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9338–9352, 2024. doi: 10.1109/TPAMI.2024.3416308. URL https://doi.org/10.1109/TPAMI.2024.3416308.

Yuhang Zhou and Zhongyun Hua. Defense without forgetting: Continual adversarial defense with anisotropic & isotropic pseudo replay. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24263–24272, 2024. doi: 10.1109/CVPR52733.2024.02290. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_Defense_without_Forgetting_Continual_Adversarial_Defense_with_Anisotropic__Isotropic_CVPR_2024_paper.html.

Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=u6TRGdzhfip.

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16443–16452, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Zi_Revisiting_Adversarial_Robustness_Distillation_Robust_Soft_Labels_Make_Student_Better_ICCV_2021_paper.html.