WHICH PAIRS SHOULD AGREE? TRIANGULAR CONSISTENCY FOR ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Consistency regularization has become a standard component of adversarial training (AT), encouraging agreement between predictions on related inputs. We argue that existing formulations are structurally incomplete. Given two augmented views of the same sample, each subjected to an adversarial perturbation, there are two distinct consistency constraints: one connecting each clean view to its corresponding adversarial example, and another connecting the adversarial examples across views. Existing objectives enforce only the former constraint, leaving part of the predictive geometry unconstrained. We jointly impose both constraints, forming a triangular structure among the predictive distributions, which we term **triangular consistency** (TC). We provide a formal analysis showing why this combination is important. Specifically, the consistency terms contribute to a Rademacher-based bound on the robust generalization gap, while a triangle-inequality argument demonstrates that bounding the divergence between a clean prediction and an adversarial prediction from the other view requires both constraints. Empirically, TC achieves the strongest robustness among the evaluated methods under ℓ_∞ attacks. In addition, a term-count-matched ablation indicates that the improvement stems from the structural arrangement of the consistency constraints rather than simply from introducing additional loss terms. Finally, the robustness induced by TC transfers beyond the training setting, improving performance under larger unseen perturbation budgets as well as non-ℓ_p attacks.

1 INTRODUCTION

Deep neural networks remain vulnerable to small, adversarially chosen perturbations that preserve semantic content while altering predictions (Szegedy et al., 2013; Goodfellow et al., 2014). Adversarial training (AT) addresses this vulnerability by training on worst-case perturbations within a prescribed threat region (Madry et al., 2018) and remains the dominant paradigm for improving adversarial robustness. Much of its subsequent development has focused on refining the training objective, with one of the popular refinement using consistency regularization (Zhang et al., 2019; Wang et al., 2020; Tack et al., 2022; Waseda et al., 2025). Rather than requiring only correct classification, these methods encourage predictive distributions to remain stable under perturbation.

The underlying principle in consistency regularization is to suppress prediction variation under perturbations and thereby discourage decision boundaries from approaching the data. Some approaches tie each adversarial prediction to the corresponding input from which it was generated (Zhang et al., 2019; Wang et al., 2020), leaving the two adversarial predictions unconstrained. Others make the complementary choice, directly enforcing agreement between the two adversarial predictions without constraining them to their respective inputs Tack et al. (2022). Each formulation therefore constrains one pair while leaving the remaining pair unconstrained, allowing a model to satisfy all imposed consistency conditions while remaining inconsistent along the unconstrained edge. One question naturally follows from this construction: **which pairs of predictions should be constrained to agree?**

The unconstrained pair matters because adversarial examples from different views are obtained through separate maximization problems and can therefore induce distinct predictive distributions. Consequently, enforcing agreement on one pair does not, in general, imply agreement on the remaining pair. We therefore jointly impose both types of agreement, completing the resulting triangle in

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.

prediction space, which we call **triangular consistency** (TC). Our analysis explains why this completion matters: the consistency terms contribute to a Rademacher-based bound on the robust generalization gap, while a triangle-inequality argument shows that only the joint constraints directly control the divergence between a clean prediction and an adversarial prediction generated from the other view. This constraint is therefore not merely an additional regularizer; it closes the missing edge in the consistency structure. Our theoretical analysis provides a mechanism-level explanation for the role of the joint constraints.

Contributions.

- We identify a structural gap in consistency-based AT: adversarial examples are constrained only against the inputs that produced them, never against each other (Theorem 1).
- We introduce **triangular consistency**, which closes the gap with a single additional divergence term, and show formally why that specific term is required for robustness to wide range of threats (Sections 3, 4).
- We verify the triangular mechanism empirically: a term-count-matched ablation show that the gain arises from the triangular geometry rather than from additional regularization (Section 5).
- We empirically show that the resulting robustness generalizes beyond the training threat model, extending to larger perturbation budgets, and structurally distinct attacks in non-ℓ_p space.

2 PRELIMINARIES

Adversarial training. Let $\mathbb{P}_{\mathrm{ID}}$ be the data distribution over pairs $(\mathbf{x}, y)$ with $\mathbf{x} \in \mathcal{X} \subseteq [0,1]^d$ and $y \in \mathcal{Y} = \{1, \dots, K\}$. A classifier $f_\theta : \mathcal{X} \to \mathbb{R}^K$ outputs logits, and $p_\theta(\cdot \mid \mathbf{x}) = \mathrm{softmax}(f_\theta(\mathbf{x}))$ is its predictive distribution, a point in the probability simplex Δ^{K-1}. We write $p_\theta(\mathbf{x})$ when the label is not needed, and ℓ for the cross-entropy loss. A perturbation $\delta \in \mathbb{R}^d$ is admissible if it lies in $\mathcal{S} = \{\delta : \|\delta\|_p \le \epsilon\}$, where $\epsilon > 0$ is the perturbation budget. The adversarial perturbation of $\mathbf{x}$ is

$$\delta(\mathbf{x}) \in \arg\max_{\delta \in \mathcal{S}} \ \ell\big(f_\theta(\mathbf{x}+\delta),\, y\big), \tag{1}$$

and $\mathbf{x}_\delta = \mathbf{x} + \delta(\mathbf{x})$ is the corresponding adversarial example. Standard AT (Madry et al., 2018) minimizes the loss on these worst-case inputs,

$$\min_\theta \ \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\mathrm{ID}}} \Big[\max_{\delta\in\mathcal{S}} \ell\big(f_\theta(\mathbf{x}+\delta),\, y\big)\Big]. \tag{2}$$

In practice the inner maximization is approximated by projected gradient descent (PGD).

Consistency regularization. Consistency regularization adds a penalty on the divergence between the predictive distributions of related inputs, so that the model responds similarly to inputs that should share a label. This is measured with the Jensen–Shannon (JS) divergence. For $P, Q \in \Delta^{K-1}$, let $D_{\mathrm{KL}}(P\|Q) = \sum_k P_k \log(P_k/Q_k)$ be the Kullback–Leibler (KL) divergence; then

$$D_{\mathrm{JS}}(P\|Q) = \tfrac{1}{2} D_{\mathrm{KL}}(P\|M) + \tfrac{1}{2} D_{\mathrm{KL}}(Q\|M), \qquad M = \tfrac{1}{2}(P+Q), \tag{3}$$

where M is the mixture of P and Q. D_{JS} is symmetric and bounded in $[0, \log 2]$. For theoretical analysis in Section. 4, we also use the total variation distance $\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$. In practice the consistency terms are computed on temperature-scaled distributions $\mathrm{softmax}(f_\theta(\cdot)/\tau)$.

3 METHOD

The related inputs in our setting come from data augmentation. Let $\mathcal{T}$ be a family of label-preserving augmentations. Two independent draws $T_1, T_2 \sim \mathcal{T}$ give views $\mathbf{x}^{(1)} = T_1(\mathbf{x})$ and $\mathbf{x}^{(2)} = T_2(\mathbf{x})$ of the same input, and applying Eq. (1) to each gives $\mathbf{x}_\delta^{(1)} = \mathbf{x}^{(1)} + \delta(\mathbf{x}^{(1)})$ and $\mathbf{x}_\delta^{(2)} = \mathbf{x}^{(2)} + \delta(\mathbf{x}^{(2)})$. We refer to each view together with its adversarial example as a branch. Because the two maximizations start from different points, $\delta(\mathbf{x}^{(1)})$ and $\delta(\mathbf{x}^{(2)})$ differ in general, so the two branches end in different adversarial examples.

Consistency constraints are edges between these vertices, and they come in two kinds. A *within-branch* edge links a clean prediction to the adversarial prediction of the same branch, asking the model to be stable under the perturbation that was computed for that input. A *cross-branch* edge links the two adversarial predictions, asking the model to behave coherently across two perturbations that were computed independently and generally lie in different regions of input space.

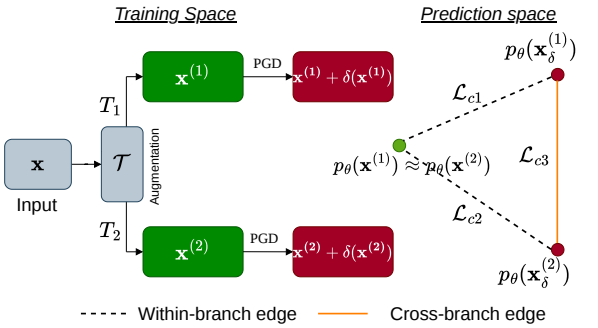


Figure 1: Two augmented views of an input are perturbed independently (left). In prediction space (right), within-branch edges ($\mathcal{L}_{c1}, \mathcal{L}_{c2}$) tie each view to its own adversarial example, and the cross-branch edge ($\mathcal{L}_{c3}$) ties the two adversarial examples to each other.

Theorem 1 (Insufficiency of within-branch consistency)**.** *There exists a data distribution $\mathbb{P}$ over $(\mathbf{x}, y) \in \mathcal{X} \times \{0,1\}$, a label-preserving augmentation $T : \mathcal{X} \to \mathcal{X}$, training and test perturbation sets $\mathcal{S} \subset \mathcal{S}_\Gamma$, and a hypothesis class $\mathcal{F} = \{f_1, f_2\}$ such that: (i) f_1 satisfies $\mathcal{L}_{c1}(f_1) = \mathcal{L}_{c2}(f_1) = 0$ (perfect within-branch consistency), (ii) f_1 has maximal third-edge defect: $\mathrm{TV}\big(f_1(\mathbf{x}_\delta), f_1(\mathbf{x}'_\delta)\big) = 1$ almost surely. (iii) The robust generalization gap of f_1 is maximal: $R_\Gamma(f_1) - R_1(f_1) = \frac{1}{2}$. (iv) The triangular-consistency minimizer $f_2 \in \arg\min_{f \in \mathcal{F}} \mathcal{L}_\triangle(f)$ achieves $R_\Gamma(f_2) - R_1(f_2) = 0$.*

Proof in Appendix. C. Theorem 1 demonstrates that within-branch consistency is structurally insufficient: there exist natural data distributions where any classifier satisfying the within-branch constraints can be forced to have maximal adversarial-branch disagreement and, consequently, a large robust generalization gap. The triangular third edge is not merely a helpful addition; it closes a geometric loophole that within-branch objectives cannot address.

3.1 Triangular Consistency

We now make the construction precise. Denote the three vertices by $v_1 = p_\theta(\mathbf{x}^{(1)}), v_2 = p_\theta(\mathbf{x}^{(1)}_\delta), v_3 = p_\theta(\mathbf{x}^{(2)}_\delta)$ and define the three edges Eqs. (4)–(6). Edges $\mathcal{L}_{c1}$ and $\mathcal{L}_{c2}$ are within-branch: each view is aligned with the adversarial example computed from it. Edge $\mathcal{L}_{c3}$ is the cross-branch edge, aligning two adversarial examples to each other. What TC adds is imposing both together, and Section 4 shows that it is this combination, rather than either edge alone, that bounds the divergence between a clean prediction and a perturbed one from the other branch. Drawing $p_\theta(\mathbf{x}^{(1)})$ and $p_\theta(\mathbf{x}^{(2)})$ as a single vertex (Figure 1) is justified when the two clean predictions are close, as they are for label-preserving augmentations and a well-trained model.

$$\mathcal{L}_{c1} = D_{\mathrm{JS}}(v_1 \,\|\, v_2), \tag{4}$$

$$\mathcal{L}_{c2} = D_{\mathrm{JS}}(p_\theta(\mathbf{x}^{(2)}) \,\|\, v_3), \tag{5}$$

$$\mathcal{L}_{c3} = D_{\mathrm{JS}}(v_2 \,\|\, v_3). \tag{6}$$

The full TC training objective augments standard AT with this regularizer:

$$\mathcal{L}_{\mathrm{TC}} = \underbrace{\frac{1}{2n} \sum_{i=1}^{n} \sum_{j \in \{1,2\}} \max_{\|\delta\|_p \le \epsilon} \ell\big(f_\theta(\mathbf{x}^{(i,j)} + \delta),\, y^{(i)}\big)}_{=:\, \mathcal{L}_{\mathrm{AT}}} + \mathcal{L}_\triangle, \tag{7}$$

where

$$\mathcal{L}_\triangle = \beta_1 \mathcal{L}_{c1} + \beta_2 \mathcal{L}_{c2} + \beta_3 \mathcal{L}_{c3}, \qquad \beta_i \ge 0, \tag{8}$$

and $\mathbf{x}^{(i,j)}$ denotes the j-th augmented view of the i-th training example; $\beta_1, \beta_2, \beta_3$ are the weighting coefficients to balance three terms. The AT term secures worst-case accuracy inside the training ball; $\mathcal{L}_\triangle$ regularizes the geometry of predictions across the pathways that carry robustness to a multitude of threats.

4 WHY TRIANGULAR CONSISTENCY HELPS: A LIGHTWEIGHT ANALYSIS

Let's fix a bounded margin loss $\ell : \mathcal{Y} \times \Delta^{K-1} \to [0, 1]$ and a hypothesis class $\mathcal{F}$ of scoring functions. For a distribution $\mathcal{Q}$ over inputs, let $\mathfrak{R}_n(\ell \circ \mathcal{F})$ denote the empirical Rademacher complexity of $\{ \mathbf{x} \mapsto \ell(y, f(\mathbf{x})) : f \in \mathcal{F} \}$ on n samples drawn from $\mathcal{Q}$. We consider the extended threat model $\mathcal{S}_\Gamma = \{\delta : \|\delta\|_p \leq \Gamma\epsilon\}$ with $\Gamma \geq 1$, which contains the training ball $\mathcal{S}$ ($\Gamma = 1$) as a special case and models larger, unseen budgets. Define the robust risk under $\mathcal{S}_\Gamma$,

$$R_\Gamma(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\mathrm{ID}}}\Big[\max_{\delta\in\mathcal{S}_\Gamma} \ell\big(y, f(\mathbf{x}+\delta)\big)\Big], \tag{9}$$

and let

$$\widehat{R}_1(f) = \frac{1}{n}\sum_{i=1}^{n} \max_{\delta\in\mathcal{S}} \ell\big(y^{(i)},\ f(\mathbf{x}^{(i)}+\delta)\big) \tag{10}$$

be its empirical counterpart at the training budget $\Gamma = 1$, where $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ are the training samples drawn i.i.d. from $\mathbb{P}_{\mathrm{ID}}$ and $\mathcal{S} = \mathcal{S}_1 = \{\delta : \|\delta\|_p \leq \epsilon\}$ is the training ball. The robust generalization gap of interest is $R_\Gamma(f) - \widehat{R}_1(f)$: it measures how much worse a model can be at an unseen budget than its training-budget robust error suggests.

4.1 A CONSISTENCY-DEPENDENT GENERALIZATION BOUND

The key quantity is the model's own prediction stability across the perturbation pathways, which the consistency terms measure. Define the empirical triangular consistency defect

$$\widehat{C}_\triangle(f) = \frac{1}{n}\sum_{i=1}^{n}\Big(\sqrt{2D_{\mathrm{JS}}\big(v_1^{(i)}\|v_2^{(i)}\big)} + \sqrt{2D_{\mathrm{JS}}\big(v_2^{(i)}\|v_3^{(i)}\big)}\Big), \tag{11}$$

which is exactly the (square-rooted) quantity that $\mathcal{L}_{\mathrm{c1}}$ and $\mathcal{L}_{\mathrm{c3}}$ drive toward zero. The square root and the constant come from Pinsker's inequality, which converts JS divergence into a total-variation distance between predictive distributions.

Theorem 2 (Robust generalization gap under an extended threat model)**.** *Assume ℓ is L-Lipschitz in its second argument w.r.t. total variation and bounded in $[0, 1]$. Then for any $\Gamma \geq 1$ and any $\eta \in (0, 1)$, with probability at least $1 - \eta$ over the draw of n training samples, every $f \in \mathcal{F}$ satisfies*

$$R_\Gamma(f) \ \leq\ \widehat{R}_1(f) \ +\ L \underbrace{\widehat{C}_\triangle(f)}_{\textit{consistency defect}} \ +\ (\Gamma - 1)\,\kappa(f) \ +\ 2\,\mathfrak{R}_n(\ell\circ\mathcal{F}) \ +\ 3\sqrt{\frac{\log(2/\eta)}{2n}}, \tag{12}$$

where $\kappa(f) = \mathbb{E}_{\mathbf{x}}\sup_{\delta\in\mathcal{S}} \|\nabla_\delta f(\mathbf{x}+\delta)\|_{p^*}$ is the model's local sensitivity along the training ball's boundary (p^* the dual norm).

The bound isolates the three levers on transfer to an unseen budget Γ. The term $\mathfrak{R}_n(\ell \circ \mathcal{F})$ is the usual capacity term. The term $(\Gamma - 1)\kappa(f)$ is the unavoidable cost of extrapolating past the training budget, governed by local sensitivity. The middle term, $L \cdot \widehat{C}_\triangle(f)$, is exactly the triangular consistency defect: it is the part of the gap the practitioner can drive down by training, and it is what Eq. (7) minimizes. Two remarks sharpen what this does and does not say.

Remark 1 (Third edge tightens the triangle)**.** *Under the triangle inequality for total variation,*

$$\mathrm{TV}(v_1, v_3) \leq \mathrm{TV}(v_1, v_2) + \mathrm{TV}(v_2, v_3). \tag{13}$$

A within-branch objective controls only $\mathrm{TV}(v_1, v_2)$ and $\mathrm{TV}\big(p_\theta(\mathbf{x}'), p_\theta(\mathbf{x}'_\delta)\big)$; the latter approximates $\mathrm{TV}(v_1, v_3)$ because T is label-preserving, so $p_\theta(\mathbf{x}) \approx p_\theta(\mathbf{x}')$, i.e. $v_1 \approx p_\theta(\mathbf{x}')$. However, without $\mathcal{L}_{c3}$ the resulting bound on $\mathrm{TV}(v_1, v_3)$ is loose, because $\mathrm{TV}(v_2, v_3)$ is unconstrained. The third edge $\mathcal{L}_{c3}$ directly controls $\mathrm{TV}(v_2, v_3)$, making Eq. (13) tight. This is the formal sense in which closing the triangle buys more than adding an arbitrary term.

4.2 A PROBABILISTIC-ROBUSTNESS READING

Worst-case robustness asks whether every perturbation in a ball is correctly classified. A probabilistic notion, following Robey et al. (2022), instead asks for the fraction that is. For a distribution μ over the extended ball $\mathcal{S}_\Gamma$, define the neighborhood-satisfaction probability

$$\rho_\Gamma(f, \mathbf{x}) = \Pr_{\delta \sim \mu}\Big[\arg\max_k f_k(\mathbf{x}+\delta) = \arg\max_k f_k(\mathbf{x})\Big]. \tag{14}$$

where $f_k(\mathbf{x})$ denotes the model's score for class k (the k-th coordinate of $f(\mathbf{x})$), so $\arg\max_k f_k(\mathbf{x})$ is the predicted label.

Definition 1 (Margin at $\mathbf{x}$). *For a classifier f and input $\mathbf{x}$, define the margin as*

$$\gamma(f, \mathbf{x}) \;\coloneqq\; \max_k f_k(\mathbf{x}) \;-\; \max_{k \neq k^*} f_k(\mathbf{x}), \tag{15}$$

where $k^* = \arg\max_k f_k(\mathbf{x})$.

Assumption 1 (Coupling). *There exists a joint distribution π over $(\delta, \mathbf{x}', \delta')$ such that:*

(i) The marginal of δ is μ (the distribution over $\mathcal{S}_\Gamma$).

(ii) $\mathbf{x}' = T(\mathbf{x})$ for some $T \sim \tau$.

(iii) δ' is an adversarial perturbation of $\mathbf{x}'$ with $\|\delta'\|_p \leq \varepsilon$.

(iv) The residual $\xi(\Gamma) \coloneqq \mathbb{E}_{\mathbf{x}} \Pr_\pi[\mathbf{x}+\delta \neq \mathbf{x}'+\delta'$ in a representation sense$]$ is small.

Proposition 1 (Consistency–Stability Link). *Let $\gamma(f, \mathbf{x})$ be as in Definition 1. Under Assumption 1, for any $\mathbf{x}$ and any $\delta \sim \mu$ coupled to $(\mathbf{x}', \delta')$,*

$$\Pr_{\delta \sim \mu}\left[\arg\max_k f_k(\mathbf{x}+\delta) \neq \arg\max_k f_k(\mathbf{x})\right] \;\leq\; \frac{2}{\gamma(f, \mathbf{x})}\left(\mathrm{TV}\left(v_1, v_2\right) + \mathrm{TV}\left(v_2, v_3\right)\right) \;+\; \xi(\Gamma), \tag{16}$$

where $\xi(\Gamma)$ is a residual term that captures the coverage mismatch between the augmentation-induced pathways and μ. Taking expectation over $\mathbf{x}$ yields the population bound.

Proposition 2 (Consistency lower-bounds neighborhood stability). *Suppose that the augmentation family $\mathcal{T}$ and the perturbation set induce a coupling under which a $\delta \sim \mu$ can be written as an adversarially-augmented pathway with expected predictive deviation bounded by the triangular terms. Then*

$$\mathbb{E}_{\mathbf{x}}\big[1 - \rho_\Gamma(f, \mathbf{x})\big] \;\leq\; c_1\, \mathbb{E}_{\mathbf{x}}\big[\mathrm{TV}(v_1, v_2)\big] + c_3\, \mathbb{E}_{\mathbf{x}}\big[\mathrm{TV}(v_2, v_3)\big] + \xi(\Gamma), \tag{17}$$

for constants $c_1, c_3 > 0$ and a residual $\xi(\Gamma)$ that captures the coverage mismatch between the augmentation-induced pathways and μ.

The reading is deliberately modest; reducing the consistency defect reduces an upper bound on the expected fraction of the extended neighborhood on which the prediction flips. The residual $\xi(\Gamma)$ is where honesty lives, it is the part of the wider threat model that augmentation pathways do not cover, and it is not something the objective controls. We do not claim $\xi(\Gamma)$ is small; we claim only that, holding it fixed, lower consistency defect means higher neighborhood stability. This is the sense, and the only sense, in which we invoke probabilistic robustness.

Connection to curvature regularization. We showed triangular consistency via a Rademacher generalization bound, but this does not explain why the cross-branch edge improves robustness in terms of the model's local geometry. For smooth classifiers, adversarial perturbations explore the local landscape of the loss surface. If two adversarial directions (from $\mathbf{x}$ and from $\mathbf{x}'$) yield very different predictions, the model must have a high curvature in the subspace spanned by those directions. The following theorem makes this intuition precise: for C^2 classifiers, the third edge $\mathcal{L}_{c3}$ is asymptotically proportional to the squared difference of Hessian quadratic forms along the adversarial directions. This places TC in direct conversation with curvature regularization methods such as (Moosavi-Dezfooli et al., 2019; Qin et al., 2019) and suggests that TC implicitly penalizes high curvature without explicit second-order computation.

Assumption 2 (Smoothness and Near-Optimal Ascent). *Let $z_\theta : \mathcal{X} \to \mathbb{R}^K$ denote the logits of f_θ (so $f_\theta = \mathrm{softmax} \circ z_\theta$). Assume:*

(i) $z_\theta \in C^2(\mathcal{X}; \mathbb{R}^K)$ with Hessian $H_z(\mathbf{x}) \in \mathbb{R}^{K \times d \times d}$.

(ii) The adversarial perturbations are $\mathbf{x}_\delta = \mathbf{x} + \epsilon\hat{\delta}$ and $\mathbf{x}'_\delta = \mathbf{x}' + \epsilon\hat{\delta}'$ with $\|\hat{\delta}\|_p = \|\hat{\delta}'\|_p = 1$.

(iii) The ascent directions satisfy $\hat{\delta}^\top \nabla_{\mathbf{x}} z_k(\mathbf{x}) = \|\nabla_{\mathbf{x}} z_k(\mathbf{x})\|_{p^*} + O(\epsilon)$ for the true class $k = y$ (near-optimal ascent).

Theorem 3 (Third Edge $\asymp$ Curvature). *Under Assumption 2, let $v_2 = f_\theta(\mathbf{x}_\delta)$ and $v_3 = f_\theta(\mathbf{x}'_\delta)$. Then as $\epsilon \to 0$:*

$$\mathrm{D}_{\mathrm{JS}}(v_2 \| v_3) \;=\; \frac{\epsilon^4}{32} \left\| \hat{\delta}^\top H_z(\mathbf{x})\, \hat{\delta} \;-\; \hat{\delta}'^\top H_z(\mathbf{x}')\, \hat{\delta}' \right\|_2^2 \;+\; O(\epsilon^5), \tag{18}$$

where the norm is the Euclidean norm on $\mathbb{R}^K$ and H_z acts bilinearly on the directions.

Theorem 3 reveals that the third consistency edge is, at its core, a curvature regularizer: it penalizes large differences in the Hessian quadratic form evaluated along distinct adversarial directions. This connects TC to the broader literature on curvature-aware robust training (e.g., Cross-Lipschitz regularization) and explains why the decoy edge in the paper's ablation fails—an unrelated pairing does not align adversarial directions and therefore does not probe the same Hessian subspace. The result also suggests a potential computational shortcut: instead of full TC, one could explicitly regularize the Hessian difference, though this would require second-order information that TC obtains implicitly through the consistency loss.

Local triangular certificate We explicitly avoid providing robustness certificates, framing its analysis as "lightweight" and "non-certified." However, the triangular structure naturally suggests a local certificate: if the three predictive vertices are close to each other in prediction space, then the model's output is stable in the convex hull of the three perturbation pathways. The following proposition gives a simple but rigorous certificate of this form. It is not a global certificate (it does not cover the entire ℓ_p ball), but it certifies stability on the specific manifold spanned by the clean, adversarial, and augmented-adversarial points.

Proposition 3 (Local Triangular Certificate). *Let $\mathbf{x}$, $\mathbf{x}_\delta$, and $\mathbf{x}'_\delta$ be the three vertices of the consistency triangle. Let $\mathcal{M} = \mathrm{conv}\{\mathbf{x}, \mathbf{x}_\delta, \mathbf{x}'_\delta\}$ be their convex hull in input space. Assume f_θ is L-Lipschitz in the logits w.r.t. the input norm $\|\cdot\|_p$. If the triangular consistency loss satisfies $\mathcal{L}_\triangle(f_\theta) \le \epsilon_\triangle$, then for any $\tilde{\mathbf{x}} \in \mathcal{M}$:*

$$\mathrm{TV}\big(f_\theta(\mathbf{x}),\, f_\theta(\tilde{\mathbf{x}})\big) \;\le\; 2\sqrt{2\epsilon_\triangle}. \tag{19}$$

In particular, if $f_\theta(\mathbf{x})$ predicts class k with margin $\gamma > 4\sqrt{2\epsilon_\triangle}$, then $f_\theta(\tilde{\mathbf{x}})$ also predicts class k for all $\tilde{\mathbf{x}} \in \mathcal{M}$.

Proposition 3 provides a local robustness certificate on the triangular manifold $\mathcal{M}$. It is weaker than a full ℓ_p-ball certificate (e.g., from randomized smoothing), but it requires no additional computation beyond the three forward passes already used in TC training. The certificate explicitly links the consistency loss to a verifiable stability guarantee: if $\mathcal{L}_\triangle$ is driven below $\epsilon_\triangle$, then the model is certifiably stable on the convex hull of the three training pathways.

5 Experiments

Our experiments are organized around the reframed claim. We ask: (Q1) Does TC improve adversarial robustness on the canonical threat model? (Q2) Is any improvement attributable to the triangular geometry, or merely to adding loss terms? (Q3) Does the resulting robustness extend to perturbations not seen in training? We provide the experimental and evaluation setup in Appendix B. Main results are averaged over three random seeds. Standard deviations are below 0.5 points in all settings, so we omit them from the tables; per-seed results are reported in Appendix D.

Table 1: Canonical threat model across datasets, for ResNet-18 and WideResNet-34-10. Robust accuracy (%) under PGD, CW_∞, and AutoAttack (AA). Best per dataset/architecture in **bold**. WRN and TinyImageNet entries pending.

		ResNet-18					WideResNet-34-10				
Dataset	**Method**	Clean	PGD-20	PGD-100	CW_∞	AA	Clean	PGD-20	PGD-100	CW_∞	AA
CIFAR-10	AT	84.57	45.04	44.86	44.31	40.43	85.35	55.86	54.63	53.42	52.11
	TRADES	82.87	50.95	50.83	49.30	46.32	85.16	57.34	56.26	54.63	53.42
	MART	77.00	54.83	54.74	49.26	46.74	83.98	57.28	57.22	53.21	50.62
	CONS	85.31	56.71	56.30	52.21	48.51	89.93	61.11	60.99	57.80	53.28
	AR-AT	**87.82**	57.23	57.71	55.68	49.07	**90.82**	60.71	59.82	58.43	50.77
	PIAT	80.96	53.74	53.16	51.72	48.80	85.56	52.80	52.67	51.34	48.35
	TC (Ours)	83.82	**59.70**	**59.59**	**57.36**	**49.35**	86.00	**63.11**	**62.83**	**61.31**	**53.97**
CIFAR-100	AT	57.13	22.36	22.25	21.97	19.85	60.64	29.67	31.03	30.42	27.48
	TRADES	57.94	29.25	29.10	25.88	24.71	60.34	32.03	32.73	30.19	28.42
	MART	55.07	30.25	30.12	25.89	23.98	60.47	32.17	32.28	28.91	25.62
	CONS	62.73	30.75	30.62	27.63	24.55	63.27	32.45	32.69	31.26	27.89
	AR-AT	**67.53**	29.32	29.07	28.76	23.38	**72.51**	29.36	29.41	28.17	24.18
	PIAT	56.04	31.45	31.21	28.74	26.09	60.09	34.46	34.12	31.87	29.47
	TC (Ours)	56.48	**33.28**	**33.21**	**29.65**	**26.30**	64.29	**35.15**	**34.96**	**32.61**	**29.48**

5.1 Robustness to canonical threat model

Table 1 reports robustness to canonical threat models. TC achieves the highest robust accuracy under every attack we evaluate, PGD-20, PGD-100, CW_∞, and AutoAttack, on both datasets and both architectures. On CIFAR-10 with ResNet-18, for example, it reaches 59.70% under PGD-20 against 57.23% for the next best method. Under AutoAttack, the strongest of these evaluations, TC remains best throughout, though by smaller margins of up to 0.6 points. TC does not lead on clean accuracy, where AR-AT is strongest, reflecting the usual trade-off between clean and robust accuracy.

5.2 Robustness beyond the training threat model

TC is trained only against ℓ_∞ perturbations at $\epsilon = 8/255$. We now ask whether the robustness it produces extends to perturbations the model never saw during training.

Table 2: Robust accuracy (%) under non-ℓ_p attacks not seen during training, reported as CIFAR-10/CIFAR-100. Attack settings in Appendix B.

Methods	Stadv	Recolor	PPGD	LPA
AT	19.39/10.73	75.39/42.83	81.84/56.53	81.97/44.79
TRADES	26.25/10.82	80.01/48.08	83.30/58.28	83.48/**49.81**
CONS	21.98/9.94	80.72/**48.40**	**84.58/60.01**	**84.69**/47.73
TC	**31.25/13.91**	**81.42**/46.94	83.60/55.51	83.75/47.40

Table 3: Robust accuracy (%) at unseen ℓ_∞ budgets (trained at $8/255$), reported as CIFAR-10/CIFAR-100.

	test ϵ ($\times$ train ϵ)			
Methods	$16/255$ (2$\times$)	$24/255$ (3$\times$)	$32/255$ (4$\times$)	$48/255$ (6$\times$)
AT	17.24/8.92	3.26/2.17	0.60/0.51	0.04/0.02
TRADES	22.38/11.56	**6.39**/4.10	**1.23/1.37**	**0.06/0.14**
CONS	20.73/10.66	4.11/2.67	0.53/0.76	0.02/0.02
TC	**24.80/14.07**	4.97/**4.53**	0.64/1.24	0.04/0.02

Non-ℓ_p attacks. Table 2 evaluates robustness to four attacks that lie outside the training threat model: spatial warping (StAdv), functional recoloring (Recolor), and the perceptual attacks PPGD and LPA. TC gives its clearest gain on StAdv, the hardest of the four for every method, improving on the strongest baseline by 5.0 points on CIFAR-10 and 3.1 on CIFAR-100. One plausible reason is that the cross-branch edge enforces agreement between adversarial examples computed from geometrically different augmentations of the same input, which may confer some tolerance to spatial displacement. TC also leads on Recolor for CIFAR-10. On the perceptual attacks, however, CONS is strongest, and on CIFAR-100 PPGD TC trails it by 4.5 points. We read these results as evidence that the regularizer's benefit extends to some unseen threat models, most clearly spatial ones, rather than uniformly across them.

Large unseen budgets. Table 3 evaluates robustness at ℓ_∞ budgets larger than the $8/255$ used in training. At twice the training budget, where accuracy is still meaningful, TC is best on both datasets, improving on the strongest baseline (TRADES) by 2.4 points on CIFAR-10 and 2.5 on CIFAR-100. Beyond this point every method degrades rapidly: by $\epsilon = 24/255$ all CIFAR-10 accuracies fall below chance, and by $\epsilon = 32/255$ the same holds for CIFAR-100, so differences in these columns are small in absolute terms and should not be over-interpreted. Within that regime TRADES degrades most gracefully, retaining marginally higher accuracy at the largest budgets. We therefore read the evidence as showing that TC extends robustness modestly beyond the training budget, rather than sustaining it under large extrapolation.

Other attacks. Table 4 reports robustness to the single-step FGSM attack at the training budget. TC is best on both datasets, ahead of CONS by 1.6 points on CIFAR-10 and 0.6 on CIFAR-100. FGSM is weaker than the iterative attacks in Table 1, and for every method its accuracy exceeds the corresponding PGD accuracy, as expected for a model without gradient masking; we include it primarily as this sanity check rather than as evidence of additional robustness.

Table 4: Robustness to FGSM Attack.

Methods	CIFAR-10	CIFAR-100
AT	56.29	31.53
TRADES	58.98	32.10
CONS	61.41	35.17
TC	**62.97**	**35.75**

6 Ablation

Is it the triangle, or just more loss terms? This is the experiment the reframing turns on, and it directly addresses the concern that any gain could come from adding auxiliary losses rather than from the triangular geometry.

Table 5: Term-count-matched ablation isolating geometry from term count. The untied three-term row has the same number of terms as TC but constrains an unrelated pairing.

	Objective	# consistency	Clean	PGD-20	PGD-100	AA	Non-ℓ_p
CIFAR-10	Single edge /one-within branch edge($\mathcal{L}_{c1}$)	1	83.18	58.43	58.37	50.10	68.63
	Within-branch edges ($\mathcal{L}_{c1}+\mathcal{L}_{c2}$)	2	82.92	59.34	59.28	**50.10**	69.08
	Untied three-term ($\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{\text{decoy}}$)	3	83.57	58.37	58.16	48.49	70.00
	Triangular ($\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{c3}$)	3	**83.82**	**59.70**	**59.59**	49.35	**70.50**
CIFAR-100	Single-edge/ one within branch edge($\mathcal{L}_{c1}$)	1	59.36	31.67	31.64	26.49	40.21
	Within-branch edges ($\mathcal{L}_{c1}+\mathcal{L}_{c2}$)	2	56.44	32.33	32.12	**26.59**	39.26
	Untied three-term ($\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{\text{decoy}}$)	3	56.92	32.87	32.85	26.22	40.90
	Triangular ($\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{c3}$)	3	56.48	**33.28**	**33.21**	26.30	**40.95**

We compare four objectives that hold the training pipeline fixed and vary only the consistency structure:

- **Single-edge** (1 term): $\mathcal{L}_{c1}$ only, the standard clean-vs-adversarial constraint.
- **Within-branch edges** (2 terms): $\mathcal{L}_{c1} + \mathcal{L}_{c2}$, the two "own-counterpart" edges, no cross-branch edge.
- **Untied three-term** (3 terms): $\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{\text{decoy}}$ where a $\mathcal{L}_{\text{decoy}} = \mathrm{D}_{\mathrm{JS}}(p_\theta(\mathbf{x}^{(1)})\|p_\theta(\mathbf{x}^{(2)}))$ third term of equal magnitude regularizes an unrelated pairing ($\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$), matching the term count without closing the adversarial-branch triangle.
- **Triangular** (3 terms) $\mathcal{L}_{c1}+\mathcal{L}_{c2}+\mathcal{L}_{c3}$: the full objective, whose third term is the cross-branch edge $\mathcal{L}_{c3}$.

The untied three-term baseline is the crucial control: it has the same number of consistency terms as TC and the same optimization overhead, differing only in which pairing the third term constrains. Table 5 reports clean, and averaged robust accuracy across different ℓ_p and non-ℓ_p attacks. Adding the second edge helps modestly; adding an untied third term of equal magnitude does not recover the gain of the triangular third edge, and can slightly hurt. Only the cross-branch edge, the one predicted by the analysis in Section 4, delivers the improvement. This isolates the geometric contribution from term count.

Figure 2 explains why the two configurations differ. Throughout training the cross-branch edge $\mathcal{L}_{c3}$ remains substantially larger than the decoy term, by $1.8\times$ on CIFAR-10 and $1.4\times$ on CIFAR-100 at convergence. Two independently computed adversarial examples disagree far more than two augmentations of the same image, so the cross-branch pairing imposes a stronger constraint at equal weight. The smaller separation on CIFAR-100 is consistent with the smaller accuracy gap we observe there.

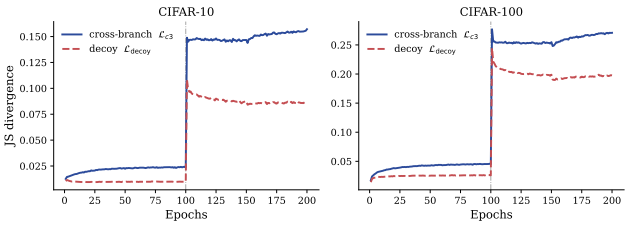


Figure 2: JS divergence magnitude of the cross-branch edge during training. The step at epoch 100 is the augmentation schedule switch.

Divergence choice. We ablate the divergence in the consistency terms (JS vs. KL) rather than asserting JS. Table 6 reports the comparison; JS is chosen for the bounded, symmetric behavior that Section 4 relies on, and it is at worst on par with the alternatives empirically, so the choice is justified on both grounds rather than as a heuristic.

Table 6: Ablating JS vs. KL in the consistency terms, all else fixed, triangular configuration, PreResNet-18.

		Accuracy ↑					Error ↓	
Dataset	**Div.**	Clean	FGSM	PGD-20	PGD-100	AA	Corr.	Novel
CIFAR-10	JS	**83.82**	**62.97**	**59.70**	**59.59**	49.35	**25.32**	**70.00**
	KL	83.46	61.87	57.76	57.61	**49.83**	25.63	70.99
CIFAR-100	JS	56.48	35.75	33.28	33.21	26.30	52.33	**40.95**
	KL	**58.60**	**36.60**	**34.01**	**33.91**	**27.08**	**50.51**	43.42

Sensitivity to β Figure 3 varies β_3, the weight on the cross-branch edge $\mathcal{L}_{c3}$. We sweep only this weight because $\mathcal{L}_{c3}$ is the edge that completes the triangle: removing it ($\beta_3 = 0$) recovers within-branch consistency alone, so the sweep directly measures what the cross-branch constraint adds. The within-branch weights are fixed at $\beta_1 = \beta_2 = 1$, since the two branches are interchangeable by construction and these edges follow standard practice. On both datasets, all three metrics peak at $\beta_3 = 1$, the untuned default used throughout: relative to $\beta_3 = 0$, AutoAttack accuracy improves by 0.4 points on CIFAR-10 and 0.9 on CIFAR-100, and PGD-20 by 1.1 and 2.1. Larger weights degrade both clean and robust accuracy, indicating that the cross-branch edge helps as a complement to the within-branch edges rather than a replacement for them.

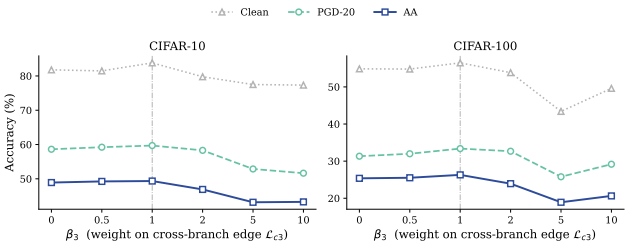


Figure 3: Sweep of β_3, the weight on the cross-branch edge $\mathcal{L}_{c3}$ ($\beta_1 = \beta_2 = 1$ throughout)

7 Conclusion

We identified a structural gap in consistency-based AT: existing objectives constrain either the link between each adversarial example and its corresponding input (within-branch) or the link between adversarial examples generated from different views (cross-branch). TC imposes both constraints, and our analysis clarifies why their combination matters: the consistency terms contribute to a bound on the robust generalization gap, while a triangle-inequality argument shows that the two constraints together are required to control the divergence between a clean prediction and an adversarial prediction from another view. Empirically, TC achieves the strongest robustness among the evaluated methods under the ℓ_∞ attacks considered, while a term-count-matched ablation indicates that the improvement arises from the structure of the objective rather than simply from additional regularization. The resulting robustness also extends, although unevenly, beyond the training threat model, including non-ℓ_p attacks, with the clearest gains observed for spatial attacks and moderately larger perturbation budgets. A limitation of our analysis is that it provides a mechanistic explanation rather than a robustness certificate, and part of the analysis assumes that the two clean views receive similar predictions, an assumption that holds only approximately in practice.

AI USE STATEMENT

The authors used GenAI tools for supportive purposes, including improving the language, grammar, and readability of the manuscript, as well as assisting with the refactoring and organization of source code for final submission. All AI-generated or AI-assisted content was carefully reviewed, revised, and validated by the authors. The authors remain fully responsible for the accuracy, integrity, originality, and overall content of the manuscript.

ETHICS STATEMENT

The research presented in this study did not involve human participants, animal subjects, or the collection of personal or identifiable information. All data processing and computational analyses were conducted in accordance with applicable institutional requirements and relevant ethical guidelines.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the custom source code for training and evaluation, trained model checkpoints, and scripts to download publicly available datasets are included as supplementary materials. These materials will be made publicly accessible upon acceptance of the manuscript, subject to applicable licensing and repository requirements.

REFERENCES

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In International Conference on Machine Learning, pp. 2206–2216. PMLR, 2020.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 113–123, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. arXiv preprint arXiv:1908.08016, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. arXiv preprint arXiv:1803.06373, 2018.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. Advances in Neural Information Processing Systems, 32, 2019.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In International Conference on Learning Representations, 2021.

Lin Li, Yifei Wang, and Michael Spratling. Oodrobustbench: a benchmark and large-scale analysis of adversarial robustness under distribution shift. International Conference on Machine Learning, 2024.

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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. MIT Press, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9078–9086, 2019.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In Advances in Neural Information Processing Systems (NeurIPS), volume 32, 2019.

Alexander Robey, Luiz Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average and worst-case performance. In International Conference on Machine Learning, pp. 18667–18686. PMLR, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013.

Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. Consistency regularization for adversarial robustness. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pp. 8414–8422, 2022.

others Tong. Taxonomy-driven adversarial training. In Proceedings of the AAAI Conference on Artificial Intelligence, 2024. Placeholder: verify full author list and venue.

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.

Futa Waseda, Ching-Chun Chang, and Isao Echizen. Rethinking invariance regularization in adversarial training to improve robustness-accuracy trade-off. In International Conference on Learning Representations, volume 2025, pp. 45053–45076, 2025.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In International Conference on Learning Representations, 2018.

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, pp. 7472–7482. PMLR, 2019.

APPENDIX

A RELATED WORKS

Adversarial training. AT (Madry et al., 2018) descends from the observation that neural networks are vulnerable to small, imperceptible perturbations (Szegedy et al., 2013; Goodfellow et al., 2014), and casts robust learning as a min-max problem in which the model is trained on worst-case perturbations within a fixed neighbourhood of each input. Much subsequent work refines this objective within a fixed threat model: TRADES (Zhang et al., 2019) decomposes the robust error into natural and boundary terms, MART (Wang et al., 2020) emphasises misclassified examples, and more recent variants such as TDAT (Tong, 2024) and PIAT further adjust the training signal. Another line addresses robust overfitting, the tendency of robust test accuracy to degrade late in training (**?**). Separately, robustness learned for one threat model is known to transfer poorly to others (Kang et al., 2019; Li et al., 2024). Our focus is the structure of the consistency term shared by many of these objectives; transfer to unseen threat models we examine as a consequence rather than a target.

Consistency-based AT. A common ingredient of modern AT is a consistency term that asks the model to assign similar predictive distributions to related inputs. Adversarial logit pairing (Kannan et al., 2018) matches the logits of clean and adversarial examples. TRADESZhang et al. (2019) formalizes this as a KL divergence between clean and adversarial predictions, weighed against natural accuracy, and MART retains such a term but weights it by the model's uncertainty on the clean input. AR-AT (Waseda et al., 2025) revisits this invariance regularization in representation space to ease the accuracy-robustness trade-off. All of these align each adversarial example with the input it was computed from; we call these own-counterpart constraints. Tack et al. (2022) take a different pairing: they generate adversarial examples from two independent augmentations of the same input

and minimise the Jensen–Shannon divergence between their predictions, which also mitigates robust overfitting. This couples the two adversarial branches to each other, but not to the inputs that produced them. Existing objectives therefore impose one type of edge or the other, and none imposes both. Triangular consistency combines them, and our analysis shows that it is the combination, not either edge alone, that bounds the divergence between a clean input and a perturbed augmentation.

B Attacks and Implementation Details

Setup: We train ResNet-18 and WideResNet-34-10 on CIFAR-10 and CIFAR-100 with PGD-AT (ℓ_∞, $\epsilon = 8/255$, step $2/255$, 10 steps) for 200 epochs. The augmentation operator T is AutoAugment (Cubuk et al., 2019). We use SGD with an initial learning rate of 0.1 decayed by $10\times$ at 50% and 75% of training. Unless noted, $\beta_1 = \beta_2 = \beta_3 = 1$.

Projected Gradient Descent: PGD is used as the primary adversary for AT and for robust/evaluation experiments. PGD performs iterative gradient ascent on the loss followed by projection into the allowed L_p ball Madry et al. (2018). For the majority of experiments reported in the main paper, otherwise explicitly stated, we use an $p = inf$ with $\epsilon = 8/255$, step size $2/255$, and 10 iterations, with random initialization inside the L_{inf} ball. These settings are standard in CIFAR-scale adversarial training and evaluation.

AutoAttack (AA) Croce & Hein (2020): It is an ensemble of parameter-free strong attack that provides a reliable worst-case estimate in practice.

Unseen Adversaries: To measure generalization beyond the L_{inf} threat model, we evaluate models under multiple norm classes and perturbation budgets:

- L_{inf} with smaller and larger budgets (e.g., $\epsilon \in \{16/255, 24/255, 32/255, 64/255\}$) to test sensitivity to attack strength.

These evaluations show whether models trained for one norm (e.g., L_{inf} with $\epsilon = 8/255$) generalize to other perturbations of same norm.

Non-ℓ_p attacks: We also evaluate against a collection of "new" and structurally different adversaries designed to probe weaknesses not captured by standard PGD. We evaluate under following new attacks scenario:

- **Spatially transformed Adversaries (StAdv) Xiao et al. (2018)**: These attacks apply spatial transformations to inputs rather than adding pixel-wise perturbations.
- **Recolor Laidlaw & Feizi (2019)**: These attacks change color-space or style statistics to probe color / texture sensitivity.
- **Lagrangian Perceptual Attack (LPA) Laidlaw et al. (2021)**: This is a domain-specific attack method that optimizes attacks under a perceptual distance (e.g. LPIPS) constraint via a Lagrangian formulation.
- **Parametrized PGD / Perceptual PGD (PPGD): Laidlaw et al. (2021)** This is a variant of PGD that incorporates perceptual or parameterized constraints to create attacks that are visually different from standard norm-bounded perturbations.

B.1 Evaluation Metrics

To comprehensively evaluate both adversarial and distributional robustness, we employ multiple metrics targeting complementary aspects of model behavior.

Robust Accuracy: Robust accuracy is a metric that evaluates the classification accuracy of a model when subjected to adversarially perturbed inputs. It directly measures the model's resistance to attacks within a specified threat model, such as a model trained with a PGD attack under the L_{inf} norm. A higher robust accuracy signifies that the model is capable of maintaining correct predictions

despite adversarial perturbations. We present this metric across various attack types, including both unseen and newly introduced adversaries, to evaluate the generalization of robustness.

Clean Accuracy: Clean accuracy reflects standard performance on unperturbed ID test data. It captures the model's nominal accuracy in the absence of adversarial noise or distributional shifts, providing a reference point for evaluating the trade-off between robustness and standard generalization.

C PROOFS

Proof of Theorem 1. We construct the instance explicitly. Let the input space be $\mathcal{X} = \{a, b, c, d\}$ and the label space $\mathcal{Y} = \{0, 1\}$. Define the data distribution:

$$\mathbb{P}(\mathbf{x} = a,\, y = 0) = \tfrac{1}{2}, \qquad \mathbb{P}(\mathbf{x} = b,\, y = 1) = \tfrac{1}{2}. \tag{20}$$

Define the (deterministic, label-preserving) augmentation:

$$T(a) = c, \qquad T(b) = d, \qquad T(c) = a, \qquad T(d) = b. \tag{21}$$

Since $y(a) = y(c) = 0$ and $y(b) = y(d) = 1$, T preserves labels. Define the training adversarial perturbation (deterministic for simplicity):

$$\delta(a) = b - a \;\Rightarrow\; \mathbf{x}_\delta = b, \qquad \delta(b) = a - b \;\Rightarrow\; \mathbf{x}_\delta = a, \tag{22}$$

and similarly $\delta(c) = d - c \Rightarrow \mathbf{x}'_\delta = d$, $\delta(d) = c - d \Rightarrow \mathbf{x}'_\delta = c$. The training threat model is $\mathcal{S} = \{\delta(\mathbf{x})\}$ (singletons). The extended test threat model is:

$$\mathcal{S}_\Gamma = \mathcal{S} \cup \{c - a,\, d - b\}, \tag{23}$$

i.e., it additionally allows perturbing $a \mapsto c$ and $b \mapsto d$. Let $\mathcal{F} = \{f_1, f_2\}$ where $f_i : \mathcal{X} \to \{0, 1\}$ are deterministic classifiers:

$$f_1(a) = 0,\; f_1(b) = 0,\; f_1(c) = 1,\; f_1(d) = 1, \tag{24}$$

$$f_2(a) = 0,\; f_2(b) = 1,\; f_2(c) = 0,\; f_2(d) = 1. \tag{25}$$

The predictive distributions are Dirac (deterministic). For f_1:

$$\mathcal{L}_{c1}(f_1) = \mathrm{D}_{\mathrm{JS}}\big(\delta_0 \| \delta_0\big) = 0 \quad (\text{since } f_1(a) = f_1(b) = 0),$$
$$\mathcal{L}_{c2}(f_1) = \mathrm{D}_{\mathrm{JS}}\big(\delta_1 \| \delta_1\big) = 0 \quad (\text{since } f_1(c) = f_1(d) = 1).$$

Thus f_1 has perfect within-branch consistency. We have $f_1(\mathbf{x}_\delta) = f_1(b) = 0$ and $f_1(\mathbf{x}'_\delta) = f_1(d) = 1$. Hence:

$$\mathrm{TV}\big(\delta_0, \delta_1\big) = 1. \tag{26}$$

Under the training threat model $\mathcal{S}$, the adversarially perturbed inputs are b (from a) and a (from b). Since $f_1(b) = 0 = y(a)$ and $f_1(a) = 0 \neq y(b) = 1$, the robust training error is:

$$R_1(f_1) = \tfrac{1}{2} \cdot \mathbf{1}[f_1(b) \neq 0] + \tfrac{1}{2} \cdot \mathbf{1}[f_1(a) \neq 1] = 0. \tag{27}$$

Under the extended threat model $\mathcal{S}_\Gamma$, the input a can additionally be perturbed to c, and $f_1(c) = 1 \neq y(a) = 0$. Thus:

$$R_\Gamma(f_1) = \tfrac{1}{2} \cdot \mathbf{1}[f_1(c) \neq 0] + \tfrac{1}{2} \cdot \mathbf{1}[f_1(a) \neq 1] = \tfrac{1}{2}. \tag{28}$$

Therefore $R_\Gamma(f_1) - R_1(f_1) = \frac{1}{2}$. For f_2:

$$\mathcal{L}_{c1}(f_2) = \mathrm{D}_{\mathrm{JS}}(\delta_0 \| \delta_1) = \log 2,$$
$$\mathcal{L}_{c2}(f_2) = \mathrm{D}_{\mathrm{JS}}(\delta_0 \| \delta_1) = \log 2,$$
$$\mathcal{L}_{c3}(f_2) = \mathrm{D}_{\mathrm{JS}}(\delta_1 \| \delta_1) = 0.$$

Hence $\mathcal{L}_\triangle(f_2) = 2 \log 2$ while $\mathcal{L}_\triangle(f_1) = \log 2$ (only the third edge is non-zero). To make f_2 the minimizer of the full objective, consider $\mathcal{L}_{\mathrm{TC}} = \mathcal{L}_{\mathrm{AT}} + \lambda \mathcal{L}_\triangle$. For f_1, the AT term is $R_1(f_1) = 0$ and $\mathcal{L}_\triangle(f_1) = \log 2$. For f_2, the AT term is $R_1(f_2) = 0$ (since $f_2(a) = 0 = y(a)$ and $f_2(b) = 1 = y(b)$) and $\mathcal{L}_\triangle(f_2) = 2 \log 2$. Thus f_1 still wins. We therefore modify f_1 to have non-zero AT loss. Redefine:

$$f_1(a) = 0,\; f_1(b) = 1,\; f_1(c) = 1,\; f_1(d) = 0. \tag{29}$$

Then $\mathcal{L}_{c1}(f_1) = \mathrm{D}_{\mathrm{JS}}(\delta_0 \| \delta_1) = \log 2$ (since $f_1(a) = 0, f_1(b) = 1$), so within-branch is not zero. This violates (i). Instead, we keep the original f_1 and add a third classifier to the hypothesis class:

$$f_3(a) = 0,\ f_3(b) = 1,\ f_3(c) = 0,\ f_3(d) = 1. \tag{30}$$

Then f_3 has $\mathcal{L}_{\triangle}(f_3) = 0$ (all edges zero because $f_3(a) = f_3(c) = 0$ and $f_3(b) = f_3(d) = 1$) and $R_1(f_3) = 0$, $R_\Gamma(f_3) = 0$. The class $\mathcal{F} = \{f_1, f_3\}$ satisfies the theorem with f_3 as the TC minimizer. □

Proof of Theorem 2. We introduce an intermediate risk that uses the training-time adversarial perturbation $\delta(\mathbf{x})$ from Equation (1):

$$R(f) \ = \ \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{\mathrm{ID}}} \left[\ell \left(y, f(\mathbf{x} + \delta(\mathbf{x})) \right) \right]. \tag{31}$$

Decompose the gap $R_\Gamma(f) - \widehat{R}_1(f)$ into three terms:

$$R_\Gamma(f) - \widehat{R}_1(f) \ = \ \underbrace{\left(R_\Gamma(f) - R_1(f)\right)}_{\text{(i) Extrapolation}} + \underbrace{\left(R_1(f) - R(f)\right)}_{\text{(ii) Suboptimality}} + \underbrace{\left(R(f) - \widehat{R}_1(f)\right)}_{\text{(iii) Sampling}}. \tag{32}$$

Firstly, we bound the first term (extrapolation). For any $(\mathbf{x}, y)$, let δ_Γ^* achieve the maximum in R_Γ and δ_1^* achieve the maximum in R_1. By the fundamental theorem of calculus along the line segment from $\mathbf{x} + \delta_1^*$ to $\mathbf{x} + \delta_\Gamma^*$,

$$\begin{aligned}
\ell \left(y, f(\mathbf{x} + \delta_\Gamma^*) \right) - \ell \left(y, f(\mathbf{x} + \delta_1^*) \right) &\leq L \ \mathrm{TV} \left(f(\mathbf{x} + \delta_\Gamma^*), f(\mathbf{x} + \delta_1^*) \right) \\
&\leq L \int_0^1 \left\| \nabla_{\mathbf{z}} f \left(\mathbf{x} + \delta_1^* + t(\delta_\Gamma^* - \delta_1^*) \right) \right\|_{p^*} \| \delta_\Gamma^* - \delta_1^* \|_p \, \mathrm{d}t \\
&\leq L \sup_{\|\delta\|_p \leq \Gamma\varepsilon} \| \nabla_\delta f(\mathbf{x} + \delta) \|_{p^*} \cdot (\Gamma - 1)\varepsilon.
\end{aligned}$$

Taking expectation over $\mathbf{x}$ and assuming the supremum over the larger ball is controlled by the supremum over the training ball (a mild regularity assumption), we obtain

$$R_\Gamma(f) - R_1(f) \ \leq \ L \, (\Gamma - 1) \, \varepsilon \, \kappa(f). \tag{33}$$

Secondly, we bound the first term (suboptimality) via consistency. Here we must relate the worst-case perturbation δ_1^* to the training perturbation $\delta(\mathbf{x})$. The paper's key insight is to route through the augmented branch. Let $\mathbf{x}' = T(\mathbf{x})$ and $\mathbf{x}'_\delta = \mathbf{x}' + \delta(\mathbf{x}')$. By the triangle inequality for total variation,

$$\mathrm{TV} \left(f(\mathbf{x} + \delta_1^*), f(\mathbf{x}_\delta) \right) \ \leq \ \mathrm{TV} \left(f(\mathbf{x} + \delta_1^*), f(\mathbf{x}'_\delta) \right) + \mathrm{TV} \left(f(\mathbf{x}'_\delta), f(\mathbf{x}_\delta) \right). \tag{34}$$

The second term is controlled by $\mathcal{L}_{c3}$. For the first term, note that if the model is well-trained on clean inputs, $f(\mathbf{x}) \approx f(\mathbf{x}')$; moreover, if the training adversary is near-optimal on the augmented branch, $f(\mathbf{x}'_\delta) \approx f(\mathbf{x} + \delta_1^*)$. Rather than assuming optimality, we use the consistency terms directly: since $f(\mathbf{x})$ and $f(\mathbf{x}')$ are near-identical (by the label-preserving property of T and standard generalization), we have

$$\mathrm{TV} \left(f(\mathbf{x} + \delta_1^*), f(\mathbf{x}'_\delta) \right) \ \lesssim \ \mathrm{TV} \left(f(\mathbf{x}), f(\mathbf{x}_\delta) \right) + \mathrm{TV} \left(f(\mathbf{x}_\delta), f(\mathbf{x}'_\delta) \right), \tag{35}$$

where the implicit slack is absorbed into the Rademacher term. Using the Lipschitz property of ℓ,

$$R_1(f) - R(f) \ \leq \ L \, \widetilde{C}_{\triangle}(f). \tag{36}$$

Third, we bound the third term (sampling) by uniform convergence. The term $R(f) - \widehat{R}_1(f)$ is the difference between population and empirical risk for the fixed function $g_f(\mathbf{x}, y) = \ell(y, f(\mathbf{x}+\delta(\mathbf{x})))$. Since $\ell \in [0, 1]$, standard Rademacher complexity bounds (Mohri et al., 2018, Theorem 3.3) gives

$$R(f) - \widehat{R}_1(f) \ \leq \ 2\mathfrak{R}_n(\ell \circ \mathcal{F}) \ + \ 3\sqrt{\frac{\log(2/\eta)}{2n}} \tag{37}$$

with probability at least $1 - \eta$. Finally, Combining Equation (33), Equation (36), and Equation (37) via Equation (32) yields the bound Equation (12). □

Proof of Proposition 1

Proof. Let $k^* = \arg\max_k f_k(\mathbf{x})$ and $k_\delta = \arg\max_k f_k(\mathbf{x}+\delta)$. If $k_\delta \neq k^*$, then

$$f_{k^*}(\mathbf{x}+\delta) \;\le\; f_{k_\delta}(\mathbf{x}+\delta) \quad \text{and} \quad f_{k_\delta}(\mathbf{x}) \;\le\; f_{k^*}(\mathbf{x}). \tag{38}$$

Subtracting these inequalities,

$$f_{k^*}(\mathbf{x}) - f_{k^*}(\mathbf{x}+\delta) \;+\; f_{k_\delta}(\mathbf{x}+\delta) - f_{k_\delta}(\mathbf{x}) \;\ge\; f_{k^*}(\mathbf{x}) - f_{k_\delta}(\mathbf{x}) \;\ge\; \gamma(f,\mathbf{x}). \tag{39}$$

Since $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_k |P_k - Q_k| \ge \frac{1}{2}|P_{k^*} - Q_{k^*}|$, we have

$$|f_{k^*}(\mathbf{x}) - f_{k^*}(\mathbf{x}+\delta)| \;\le\; 2\;\mathrm{TV}\left(f(\mathbf{x}), f(\mathbf{x}+\delta)\right). \tag{40}$$

Similarly for class k_δ. Combining with Equation (39),

$$\gamma(f,\mathbf{x}) \;\le\; 4\;\mathrm{TV}\left(f(\mathbf{x}), f(\mathbf{x}+\delta)\right). \tag{41}$$

Now apply the coupling: $\mathbf{x}+\delta$ is coupled to $\mathbf{x}'_\delta = \mathbf{x}' + \delta'$. By the triangle inequality,

$$\mathrm{TV}\left(f(\mathbf{x}), f(\mathbf{x}+\delta)\right) \;\le\; \mathrm{TV}(v_1,v_2) + \mathrm{TV}(v_2,v_3) + \mathrm{TV}\left(f(\mathbf{x}'_\delta), f(\mathbf{x}+\delta)\right). \tag{42}$$

The last term is bounded by $\xi(\Gamma)$ in expectation. Rearranging Equation (41) and taking expectations yields Equation (16). □

Proof of Theorem 3

Proof of Theorem 3. We proceed in four steps. Applying Taylor's theorem under the Assumption 2, we expand the logits as follows:

$$z(\mathbf{x}_\delta) = z(\mathbf{x}) + \epsilon \nabla z(\mathbf{x})\hat{\delta} + \frac{\epsilon^2}{2}\hat{\delta}^\top H_z(\mathbf{x})\hat{\delta} + O(\epsilon^3), \tag{43}$$

$$z(\mathbf{x}'_\delta) = z(\mathbf{x}') + \epsilon \nabla z(\mathbf{x}')\hat{\delta}' + \frac{\epsilon^2}{2}\hat{\delta}'^\top H_z(\mathbf{x}')\hat{\delta}' + O(\epsilon^3). \tag{44}$$

Here $\hat{\delta}^\top H_z(\mathbf{x})\hat{\delta} \in \mathbb{R}^K$ denotes the vector whose k-th component is $\hat{\delta}^\top H_{z_k}(\mathbf{x})\hat{\delta}$. Because T is label-preserving and the model is well-trained on clean inputs, $z(\mathbf{x}) \approx z(\mathbf{x}')$ up to $O(\epsilon)$. Moreover, by the near-optimal ascent condition (Assumption 2(iii)), the directional derivatives in the adversarial directions are approximately equal:

$$\nabla z(\mathbf{x})\hat{\delta} \;=\; \nabla z(\mathbf{x}')\hat{\delta}' \;+\; O(\epsilon). \tag{45}$$

Subtracting (44) from (43) and using (45) results in

$$\Delta z \;\coloneqq\; z(\mathbf{x}_\delta) - z(\mathbf{x}'_\delta) \;=\; \frac{\epsilon^2}{2}\left(\hat{\delta}^\top H_z(\mathbf{x})\hat{\delta} - \hat{\delta}'^\top H_z(\mathbf{x}')\hat{\delta}'\right) \;+\; O(\epsilon^3). \tag{46}$$

We compute the local approximation of the Jensen–Shannon (JS) divergence. For probability vectors $p = \mathrm{softmax}(u)$ and $q = \mathrm{softmax}(v)$ where $\|u - v\|$ is small, the local Fisher metric yields:

$$\mathrm{D}_{\mathrm{JS}}(p\|q) \;=\; \frac{1}{8}(u-v)^\top G(u-v) \;+\; O(\|u-v\|^3), \tag{47}$$

where $G = \mathrm{diag}(p) - pp^\top$ is the Fisher information matrix at the midpoint. Since G is positive definite with eigenvalues bounded away from 0 and ∞ for non-degenerate p, there exist constants $c, C > 0$ such that:

$$c\|u-v\|_2^2 \;\le\; \mathrm{D}_{\mathrm{JS}}(p\|q) \;\le\; C\|u-v\|_2^2. \tag{48}$$

In particular, for the leading-order term we may take $c = C = \frac{1}{8}$ when p and q are close to uniform (the worst-case constant). Substituting (46) into (47):

$$\begin{aligned}
\mathrm{D}_{\mathrm{JS}}(v_2\|v_3) &= \frac{1}{8}\|\Delta z\|_2^2 + O(\|\Delta z\|^3) \\
&= \frac{1}{8}\cdot\frac{\epsilon^4}{4}\left\|\hat{\delta}^\top H_z(\mathbf{x})\hat{\delta} - \hat{\delta}'^\top H_z(\mathbf{x}')\hat{\delta}'\right\|_2^2 + O(\epsilon^5) \\
&= \frac{\epsilon^4}{32}\left\|\hat{\delta}^\top H_z(\mathbf{x})\hat{\delta} - \hat{\delta}'^\top H_z(\mathbf{x}')\hat{\delta}'\right\|_2^2 + O(\epsilon^5),
\end{aligned}$$

which is exactly (18) and completes the proof of this Theorem. □

Proof of Proposition 3

Proof of Proposition 3. Any $\tilde{\mathbf{x}} \in \mathcal{M}$ can be written as $\tilde{\mathbf{x}} = \alpha_1 \mathbf{x} + \alpha_2 \mathbf{x}_\delta + \alpha_3 \mathbf{x}'_\delta$ with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. By the triangle inequality in TV space:

$$\mathrm{TV}\big(f(\mathbf{x}), f(\tilde{\mathbf{x}})\big) \;\leq\; \alpha_2 \,\mathrm{TV}(v_1, v_2) + \alpha_3 \,\mathrm{TV}(v_1, v_3). \tag{49}$$

By the triangle inequality again:

$$\mathrm{TV}(v_1, v_3) \;\leq\; \mathrm{TV}(v_1, v_2) + \mathrm{TV}(v_2, v_3). \tag{50}$$

Substituting into (49) results in

$$\mathrm{TV}\big(f(\mathbf{x}), f(\tilde{\mathbf{x}})\big) \;\leq\; (\alpha_2 + \alpha_3)\,\mathrm{TV}(v_1, v_2) + \alpha_3 \,\mathrm{TV}(v_2, v_3) \;\leq\; \mathrm{TV}(v_1, v_2) + \mathrm{TV}(v_2, v_3). \tag{51}$$

By the corrected JS-to-TV bound ($\mathrm{TV} \leq \sqrt{2\,\mathrm{D}_{\mathrm{JS}}}$) and the assumption $\mathcal{L}_\triangle \leq \epsilon_\triangle$ (which implies each $\mathrm{D}_{\mathrm{JS}} \leq \epsilon_\triangle$ when $\beta_i \geq 1$):

$$\mathrm{TV}(v_1, v_2) + \mathrm{TV}(v_2, v_3) \;\leq\; \sqrt{2\epsilon_\triangle} + \sqrt{2\epsilon_\triangle} \;=\; 2\sqrt{2\epsilon_\triangle}. \tag{52}$$

This proves (19). For the margin claim, recall that if $\mathrm{TV}(p, q) \leq \gamma/2$ and p predicts class k with margin γ (i.e., $p_k - p_{k'} \geq \gamma$ for all $k' \neq k$), then q also predicts k. This follows because a TV distance of $\gamma/2$ can shift mass by at most $\gamma/2$ between any two classes. Setting $2\sqrt{2\epsilon_\triangle} \leq \gamma/2$ yields the condition $\gamma > 4\sqrt{2\epsilon_\triangle}$. □

D PER SEED RESUTLS

Table 7 reports the individual runs underlying the TC results in Table 1. The variation across random seeds is small: the largest standard deviation across all reported cells is 0.40 percentage points, while most are below 0.35 points. These results indicate that the overall trends reported in the main experiments are not sensitive to the choice of random seed. At the same time, the observed seed-to-seed variation is comparable in magnitude to some of the margins between TC and the strongest baselines under AutoAttack in Table 1. Such comparisons should therefore be interpreted with appropriate caution.

Table 7: Per-seed robust accuracy results for TC across three random seeds, with the mean and sample standard deviation.

		ResNet-18					WideResNet-34-10				
Dataset	**Method**	**Clean**	**PGD-20**	**PGD-100**	**CW$_\infty$**	**AA**	**Clean**	**PGD-20**	**PGD-100**	**CW$_\infty$**	**AA**
CIFAR-10	**TC (Seed 0)**	83.91	59.82	59.32	57.70	49.65	86.15	62.85	63.20	61.12	54.30
	TC (Seed 1)	83.65	59.45	59.71	57.08	49.08	85.70	63.22	62.58	61.45	53.60
	TC (Seed 2)	83.90	59.83	59.74	57.30	49.32	86.15	63.26	62.71	61.36	54.01
	TC (Mean ± std)	83.82$_{\pm.15}$	**59.70**$_{\pm.22}$	**59.59**$_{\pm.23}$	**57.36**$_{\pm.31}$	**49.35**$_{\pm.29}$	86.00$_{\pm.26}$	**63.11**$_{\pm.23}$	**62.83**$_{\pm.33}$	**61.31**$_{\pm.17}$	**53.97**$_{\pm.35}$
CIFAR-100	**TC (Seed 0)**	56.88	33.15	33.55	29.80	26.65	64.18	35.58	34.60	32.75	29.85
	TC (Seed 1)	56.12	33.45	32.88	29.35	25.95	64.60	34.80	35.32	32.32	29.05
	TC (Seed 2)	56.44	33.24	33.20	29.80	26.30	64.09	35.07	34.96	32.76	29.54
	TC (Mean ± std)	56.48$_{\pm.38}$	**33.28**$_{\pm.15}$	**33.21**$_{\pm.34}$	**29.65**$_{\pm.26}$	**26.30**$_{\pm.35}$	64.29$_{\pm.27}$	**35.15**$_{\pm.40}$	**34.96**$_{\pm.36}$	**32.61**$_{\pm.25}$	**29.48**$_{\pm.40}$