DAMPED FILTERING FOR REPRESENTATION CONDITIONED ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adversarial training typically applies a single ambient geometry to all local input directions, while manifold-aware defenses often require an unavailable hard tangent–normal decomposition. We propose **Representation-Conditioned Complementary Damped-Filter Regularization (R-CDFR)**, which uses the singular spectrum of the classifier's representation Jacobian to allocate regularization according to local representation gain. Complementary damped filters define a high-gain pass branch and a low-gain residual branch. Although the exact Jacobian-null component of the input loss gradient vanishes, finite damping yields a nontrivial row-space residual that emphasizes low-gain directions. R-CDFR combines spectrum-wide inverse-metric slope control, concentrated residual-gradient suppression, and finite-step predictive consistency along pass-filtered perturbations. All operators are computed matrix-free using Jacobian–vector and vector–Jacobian products. We derive a local loss-growth bound for the induced perturbation family and relate the three regularizers to complementary local and finite-step quantities. Extensive experiments show consistent robustness gains across datasets, training recipes, convolutional and transformer backbones, and analytic tasks with known directional geometry.

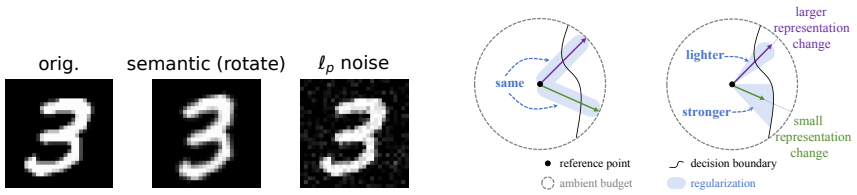
1 INTRODUCTION

Adversarial robustness seeks to stabilize predictions against harmful perturbations while preserving meaningful data variation. Most methods formulate this objective as local invariance within a fixed ambient norm ball (Madry et al., 2018; Zhang et al., 2019), which applies the same notion of perturbation size to every input direction. Yet learned representations can respond very differently across directions, so perturbations with similar ambient norms may produce substantially different feature and prediction changes. This mismatch is especially relevant in high-dimensional input spaces, where meaningful data variation occupies only a small subset of directions and the remaining directions are weakly constrained by observations. As a result, isotropic objectives cannot precisely allocate robustness according to representation response. This directional mismatch is closely related to non-robust features and the accuracy–robustness trade-off (Ilyas et al., 2019; Tsipras et al., 2019; Zhang et al., 2025), motivating manifold-aware and direction-aware analyses of adversarial risk (Zhang et al., 2022b; 2024; Xiao et al., 2025).

Turning this directional perspective into a practical objective remains difficult because the relevant structure is unknown on natural images and representation responses vary continuously across directions. Figure 1(a) contrasts meaningful local variation with ambient drift, while Figure 1(b) shows that direction-agnostic training applies uniform regularization even when weak-gain directions remain influential for prediction. This motivates a model-adaptive objective that allocates stronger regularization to low-gain, prediction-relevant directions while preserving smooth variation along high-gain directions, as illustrated in Figure 1(c).

We therefore pursue *representation-conditioned anisotropic robustness*, where the learned representation determines how regularization is distributed across local input directions. Our central idea is to preserve smooth high-gain variation while imposing stronger regularization on low-gain directions that remain influential for prediction. We operationalize this idea through the input Jacobian of the learned representation, whose singular spectrum continuously quantifies how strongly each local direction changes the representation and thereby connects ambient regularizers that aggregate across

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.



(a) Meaningful variation and ambient drift. (b) Isotropic. (c) Anisotropic.

Figure 1: Directional motivation. Meaningful local variation and weakly supported ambient drift can induce substantially different changes in a learned representation. Isotropic training uses a common ambient geometry, whereas representation-conditioned filtering allocates stronger control to low-gain directions while maintaining smooth behavior along high-gain variation.

directions with methods based on discrete directional subspaces. Building on this spectrum, we propose **Representation-Conditioned Complementary Damped-Filter Regularization (R-CDFR)** which constructs complementary damped filters with a pass branch that retains high-gain variation and a residual branch that emphasizes low-gain gradient components. R-CDFR combines an inverse-metric loss-slope penalty for spectrum-wide gain-aware control, a residual-filtered gradient penalty that concentrates suppression on the lowest-gain components, and a finite-step consistency penalty that stabilizes the predictive distribution along pass-filtered perturbations. Together, these terms integrate infinitesimal spectral regularization and finite-step predictive stability within a single differentiable objective. Damping yields stable, continuously varying operators for the rank-deficient Jacobians common in image models, while all filter and metric operations are computed matrix-free through Jacobian–vector and vector–Jacobian products. Consequently, R-CDFR provides finer directional control than ambient regularization, avoids the brittle rank selection of hard subspace partitions, and remains readily applicable to modern neural representations.

Our analysis bounds local loss growth under the damped representation-filtered perturbation family and connects the spectral-slope and residual penalties to the resulting first-order terms, while the consistency term provides complementary finite-step predictive control. Experiments on CIFAR-10, CIFAR-100, and TinyImageNet show consistent robustness gains under matched protocols and a strong modern training recipe, with similar improvements across convolutional and transformer backbones. Known-geometry tasks, ablations, and residual diagnostics further support the proposed spectral allocation mechanism. Our contributions are summarized as follows:

- We formulate representation-conditioned anisotropic robustness as continuous regularization over the local Jacobian gain spectrum, connecting ambient sensitivity control with methods based on discrete directional subspaces.
- We propose R-CDFR, a fully differentiable objective that combines spectrum-wide inverse-metric slope control, concentrated damped residual suppression, and finite-step pass-filtered consistency using the classifier's learned representation.
- We derive a local loss-growth bound for the finite-damping perturbation family, identify its pass and residual first-order coefficients, and relate the three training terms to complementary parts of this local objective.
- We provide matched robustness results, a strong-recipe add-on study, convolutional and transformer evaluations, spectrum-resolved mechanism controls, and directional experiments under analytic ground-truth geometry.

2 Related Work

Adversarial training and evaluation. Adversarial robustness is commonly pursued through PGD training, TRADES, MART, and adversarial weight perturbation (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2020; Wu et al., 2020). Efficient variants reduce the cost of inner maximization (Shafahi et al., 2019; Wong et al., 2020), while augmentation and optimization choices can substantially affect the clean–robust operating point (Wang et al., 2023; Amini et al., 2024; Bai et al., 2024; Yun et al., 2019). AutoAttack and RobustBench provide standardized evaluations (Croce &

Hein, 2020; Croce et al., 2021), and existing trade-off results caution against interpreting empirical gains as eliminating the general accuracy–robustness tension (Tsipras et al., 2019; Ilyas et al., 2019; Dohmatob, 2019). We use matched protocols and additionally evaluate R-CDFR on a stronger TRADES-AWP recipe.

Lipschitz, Jacobian, and consistency regularization. Sensitivity-based defenses constrain orthogonality, Jacobian norms, input-loss gradients, or Lipschitz margins (Cisse et al., 2017; Miyato et al., 2018; Jakubovitz & Giryes, 2018; Chan et al., 2020; Finlay et al., 2018b;a; Sokolić et al., 2017; Hein & Andriushchenko, 2017; Fazlyab et al., 2019; Tsuzuku et al., 2018). Related work studies these controls in structured, sparse, and certified settings (Muthukumar & Sulam, 2023; Zhang et al., 2022a; Zühlke & Kudenko, 2025; Chen et al., 2026). Virtual adversarial training and consistency regularization instead stabilize predictions over local perturbations (Miyato et al., 2019; Tack et al., 2022), with similar sensitivity controls also used for adversarial transfer (Wu et al., 2024; Chen et al., 2025). These approaches generally optimize ambient or aggregate sensitivity, whereas R-CDFR conditions first-order and finite-step control on the continuous singular spectrum of an internal representation.

Manifold and directional robustness. The manifold hypothesis motivates robustness objectives that exploit local geometric structure (Belkin et al., 2006). Tangent-normal regularization, manifold adversarial learning, tangent-guided training, embedded-manifold gradient control, and robust feature learning introduce explicit directional or representation-aware structure (Yu et al., 2019; Zhang et al., 2018; Yi et al., 2024; Li et al., 2020; Zhang et al., 2025). Complementary analyses decompose adversarial risk into in-manifold and normal components (Zhang et al., 2022b; 2024; Xiao et al., 2025), while projection and purification methods model the manifold more explicitly (Samangouei et al., 2018; Stutz et al., 2019; Li et al., 2025). Low-dimensional structure alone does not prevent ambient attacks (Pfrommer et al., 2023). R-CDFR builds on directional regularization by replacing a binary tangent–normal partition with complementary damped responses over the full representation-gain spectrum and combining the resulting first-order controls with finite-step predictive consistency.

3 Preliminaries

3.1 Robust classification and notation

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the input space, where d is the input dimension, and let $\mathcal{Y} = \{1, \dots, K\}$ denote the label set, where K is the number of classes. We write the classifier as $f_\theta = c_\phi \circ h_\psi$, where $h_\psi : \mathcal{X} \to \mathbb{R}^m$ is an m-dimensional internal representation, $c_\phi : \mathbb{R}^m \to \mathbb{R}^K$ is the classification head, and $\theta = (\psi, \phi)$ collects all model parameters. For a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$, define

$$\ell(x, y) := \ell_{\mathrm{ce}}(f_\theta(x), y), \qquad g := \nabla_x \ell(x, y), \qquad J := J_h(x) = \frac{\partial h_\psi(x)}{\partial x} \in \mathbb{R}^{m \times d}, \tag{1}$$

where ℓ_{ce} is the cross-entropy loss, $g \in \mathbb{R}^d$ is the input loss gradient, and J is the representation Jacobian at x.

Given a data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, standard adversarial training minimizes

$$\mathcal{L}_{\mathrm{adv}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_p \le \epsilon} \ell(x + \delta, y) \right], \tag{2}$$

where $\delta \in \mathbb{R}^d$ is an input perturbation, $\|\cdot\|_p$ is the ℓ_p norm, and $\epsilon > 0$ is the perturbation budget.

4 R-CDFR: Damped Residual Filtering

4.1 Gain-conditioned complementary filters

Recall that $J = J_h(x) \in \mathbb{R}^{m \times d}$ is the representation Jacobian, and let $G = J^\top J$ denote the corresponding local Gram operator. For a nonzero input direction $v \in \mathbb{R}^d$, differentiability of h_ψ at x gives

$$h_\psi(x + tv) = h_\psi(x) + tJv + o(|t|), \qquad t \to 0.$$

Thus, tJv is the first-order change in the representation induced by the input displacement tv. We define

$$\rho_x(v) := \frac{\|Jv\|_2}{\|v\|_2}$$

as the **directional representation gain**, which measures the magnitude of this first-order representation change per unit input displacement. Let

$$J = U_r \Sigma V_r^\top, \qquad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r), \tag{3}$$

be the compact singular value decomposition, where $r = \mathrm{rank}(J)$, $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{d\times r}$ have orthonormal columns, and $\sigma_i > 0$ are the nonzero singular values. The columns u_i of U_r and v_i of V_r lie in the representation and input spaces, respectively. Since $Jv_i = \sigma_i u_i$, each input-space singular direction v_i has gain $\rho_x(v_i) = \sigma_i$. We refer to $\{\sigma_i\}_{i=1}^r$ as the **local representation-gain spectrum**.

The central insight is to use this spectrum as a continuous weighting variable. For this purpose, we define the **damped Gram operator** A_λ, the **representation-pass filter** P_λ, and the **residual filter** R_λ as

$$A_\lambda := G + \lambda I_d, \qquad P_\lambda := J^\top (JJ^\top + \lambda I_m)^{-1} J, \qquad R_\lambda := I_d - P_\lambda, \qquad \lambda > 0, \tag{4}$$

where I_d and I_m are identity matrices in the input and representation spaces, respectively, and λ is the damping parameter. Because $R_\lambda = I_d - P_\lambda$, the filters are **complementary**: every $v \in \mathbb{R}^d$ satisfies $v = P_\lambda v + R_\lambda v$. Here the complementarity means that the two filtered components recover the original vector when added.

A straightforward calculation gives the action of the filters along each nonzero singular direction:

$$P_\lambda v_i = \frac{\sigma_i^2}{\sigma_i^2 + \lambda} v_i, \qquad R_\lambda v_i = \frac{\lambda}{\sigma_i^2 + \lambda} v_i. \tag{5}$$

Note that the pass weight increases with the directional gain σ_i, whereas the residual weight decreases with it. Thus, the pass filter retains high-gain directions more strongly, while the residual filter assigns larger weights to low-gain directions. The two weights sum to one, yielding a smooth **gain-conditioned split** without explicit rank selection. For finite λ, the two filters can retain overlapping parts of the same nonzero-gain direction, so the split is soft rather than an orthogonal decomposition.

The use of the classifier's own representation also determines how the filters act on the input loss gradient. Let $q := \nabla_h \ell_{\mathrm{ce}}(c_\phi(h_\psi(x)), y) \in \mathbb{R}^m$ be the loss gradient with respect to the representation. The chain rule gives $g = J^\top q \in \mathrm{Im}(J^\top)$, where $\mathrm{Im}(J^\top)$ is the row space of J. Hence, g has no component in $\mathrm{Ker}(J)$. Writing $a_i := v_i^\top g$ for the coefficient of g along v_i, we obtain the **low-gain spectral residual**

$$R_\lambda g = \sum_{i=1}^{r} \frac{\lambda}{\sigma_i^2 + \lambda} a_i v_i. \tag{6}$$

The residual weight is larger for smaller directional gains, so low-gain row-space components are retained more strongly. Thus, although the exact Jacobian-null component of g is zero, finite damping produces a nontrivial residual that remains informative within $\mathrm{Im}(J^\top)$. R-CDFR uses R_λ to concentrate regularization on these components and uses P_λ to construct finite-step perturbations for predictive consistency.

4.2 Representation-conditioned training objective

R-CDFR augments a base training objective $\mathcal{L}_{\mathrm{base}}(\theta)$ with an inverse-metric spectral-slope regularizer, a damped residual-gradient penalty, and a finite-step pass-filtered consistency term.

Inverse-metric spectral-slope regularizer. The damped metric A_λ measures the size of local input displacements according to their representation gain. Its dual norm therefore characterizes the worst-case first-order loss slope under an A_λ-bounded displacement:

$$\sup_{\|\delta\|_{A_\lambda} \le 1} g^\top \delta = \|g\|_{A_\lambda^{-1}}, \qquad \|g\|_{A_\lambda^{-1}} := \sqrt{g^\top A_\lambda^{-1} g}.$$

Since $g \in \mathrm{Im}(J^\top)$ and $A_\lambda v_i = (\sigma_i^2 + \lambda)v_i$, its spectral form is

$$\|g\|^2_{A_\lambda^{-1}} = \sum_{i=1}^{r} \frac{a_i^2}{\sigma_i^2 + \lambda}. \tag{7}$$

R-CDFR uses its expected value as the inverse-metric spectral-slope regularizer

$$\mathcal{L}_{\mathrm{spec}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\|g\|_{A_\lambda^{-1}} \right]. \tag{8}$$

Equation (7) assigns larger penalties to lower-gain gradient components, with λ bounding the weight near zero gain. Moreover, since $\|P_\lambda g\|_{A_\lambda^{-1}} \le \|g\|_{A_\lambda^{-1}}$, $\mathcal{L}_{\mathrm{spec}}$ provides a spectrum-wide upper surrogate for the pass-filtered first-order loss growth analyzed in Section 5.

Damped residual-gradient penalty. To directly penalize the low-gain component isolated by the residual filter, we define

$$\mathcal{L}_{\mathrm{res}} = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\|R_\lambda g\|_2^2 \right]. \tag{9}$$

For a single example, this penalty has the spectral form

$$\|R_\lambda g\|_2^2 = \sum_{i=1}^{r} \frac{\lambda^2}{(\sigma_i^2 + \lambda)^2} a_i^2. \tag{10}$$

The coefficient $\lambda^2/(\sigma_i^2 + \lambda)^2$ decreases more rapidly with σ_i than the weighting in Eq. (7). Thus, $\mathcal{L}_{\mathrm{spec}}$ provides broad gain-aware control, while $\mathcal{L}_{\mathrm{res}}$ adds a sharper penalty near the low-gain end of the spectrum.

Finite-step pass-filtered consistency. The two gradient penalties control infinitesimal, label-dependent loss variation. To additionally regularize predictive behavior over finite displacements, we construct a stochastic perturbation along directions retained by the pass filter. Specifically, we sample an isotropic direction $u \sim \mathcal{N}(0, I_d)$ and apply P_λ to reweight its components according to representation gain. If $u = \sum_{i=1}^{r} b_i v_i + u_0$ with $u_0 \in \mathrm{Ker}(J)$, then

$$P_\lambda u = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} b_i v_i,$$

so higher-gain components are retained more strongly, while exact zero-gain components are removed. We then normalize the filtered direction to control the Euclidean displacement magnitude. Let $p_\theta(x) := \mathrm{softmax}(f_\theta(x)) \in \mathbb{R}^K$ and define

$$\delta_P = \varepsilon_P \frac{P_\lambda u}{\|P_\lambda u\|_2 + \eta}, \qquad \mathcal{L}_{\mathrm{cons}} = \mathbb{E}_{\substack{(x,y)\sim\mathcal{D} \\ u\sim\mathcal{N}(0,I_d)}} \left[\mathrm{KL}\big(p_\theta(x) \,\|\, p_\theta(x+\delta_P)\big) \right], \tag{11}$$

where $\varepsilon_P > 0$ controls the displacement magnitude, $\eta > 0$ prevents division by zero, and KL denotes Kullback–Leibler divergence. This construction yields a finite pass-filtered perturbation with $\|\delta_P\|_2 < \varepsilon_P$. Unlike the two gradient penalties, $\mathcal{L}_{\mathrm{cons}}$ compares the full predictive distribution after a finite displacement and therefore captures class-probability changes beyond the scalar first-order cross-entropy slope.

Overall objective. Combining the three regularizers, R-CDFR minimizes

$$\min_\theta \quad \mathcal{L}_{\mathrm{base}}(\theta) + \alpha\mathcal{L}_{\mathrm{spec}}(\theta) + \beta\mathcal{L}_{\mathrm{res}}(\theta) + \gamma\mathcal{L}_{\mathrm{cons}}(\theta), \tag{12}$$

where $\alpha, \beta, \gamma \ge 0$ weight the inverse-metric spectral-slope, damped residual-gradient, and finite-step pass-filtered consistency terms, respectively. The complete training procedure is provided in Algorithm 1 in Appendix B.

Matrix-free computation. All Jacobian-dependent quantities are computed without explicitly forming J. Jacobian–vector products Jv and vector–Jacobian products $J^\top z$ provide the implicit matrix–vector operations required by conjugate gradients. In particular, if w solves $A_\lambda w = g$, then $\|g\|_{A_\lambda^{-1}} = \sqrt{g^\top w}$ and $R_\lambda g = \lambda w$, so $\mathcal{L}_{\mathrm{spec}}$ and $\mathcal{L}_{\mathrm{res}}$ reuse the same linear solve. Likewise, $P_\lambda u$ is obtained by solving $(JJ^\top + \lambda I_m)t = Ju$ and returning $J^\top t$. The complete algorithm and solver details are provided in Appendix B.

5 LOCAL INTERPRETATION OF THE FILTERED OBJECTIVE

We analyze local loss growth for perturbations generated by the same complementary filters used in the objective. Define the **representation-filtered perturbation family**

$$\Delta_\lambda(x;\tau_P,\varepsilon_R) := \left\{P_\lambda u_P + R_\lambda u_R : \|u_P\|_{A_\lambda} \le \tau_P,\ \|u_R\|_2 \le \varepsilon_R\right\}, \tag{13}$$

where $u_P, u_R \in \mathbb{R}^d$ are auxiliary input-space vectors, $\|u_P\|_{A_\lambda} := \sqrt{u_P^\top A_\lambda u_P}$, and τ_P and ε_R control the pass-filtered and residual-filtered components. The metric and both filters are evaluated at the reference point x and held fixed within this local perturbation family. For later use, we define the corresponding first-order coefficients $S_P(x) := \|P_\lambda g\|_{A_\lambda^{-1}}$, $S_R(x) := \|R_\lambda g\|_2$.

Proposition 5.1 (Local loss growth over the filtered perturbation family)**.** *Assume that the input gradient of $\ell(\cdot, y)$ is H-Lipschitz along every segment from x to $x+\delta$ with $\delta \in \Delta_\lambda(x;\tau_P,\varepsilon_R)$. Then*

$$\sup_{\delta\in\Delta_\lambda(x;\tau_P,\varepsilon_R)} \ell(x+\delta,y) \le \ell(x,y) + \tau_P S_P(x) + \varepsilon_R S_R(x) + \frac{H}{2}\left(\frac{\tau_P}{\sqrt{\lambda}} + \varepsilon_R\right)^2. \tag{14}$$

The proposition closes the link to the objective: $\mathbb{E}[S_P(x)] \le \mathcal{L}_{\text{spec}}$ because $\|P_\lambda g\|_{A_\lambda^{-1}} \le \|g\|_{A_\lambda^{-1}}$, while Jensen's inequality gives $\mathbb{E}[S_R(x)] \le \sqrt{\mathcal{L}_{\text{res}}}$. The first term supplies broad spectral control and the second adds low-gain concentration. The finite-step KL term is complementary to this first-order expansion and motivates the third diagnostic, G_P. The experiments therefore evaluate S_P, S_R, and G_P alongside robust accuracy. The proof appears in Appendix A.

6 EXPERIMENTS

Research questions. The experiments follow the objective and the local interpretation above. **RQ-A** Does R-CDFR improve standard adversarial robustness under matched training and evaluation? **RQ-B** Does it reduce the pass and residual first-order coefficients and the finite-step predictive divergence targeted by the three regularizers? **RQ-C** Is the suppression concentrated in low-gain directions, and does it translate into larger decision-changing displacements under residual-filtered and known-geometry probes? **RQ-D** Do the component ablations and parameter controls recover the distinct roles assigned to $\mathcal{L}_{\text{spec}}$, $\mathcal{L}_{\text{res}}$, and $\mathcal{L}_{\text{cons}}$?

6.1 EXPERIMENTAL SETTING

We evaluate CIFAR-10, CIFAR-100, and TinyImageNet under the standard ℓ_∞ threat model with $\epsilon = 8/255$, reporting PGD-20 and the official default AutoAttack ensemble. CIFAR-10 and CIFAR-100 results are averaged over five independent seeds, and TinyImageNet results over three seeds. All attacks operate on raw $[0,1]$ inputs, with channel normalization included inside the attacked model wrapper. Unless stated otherwise, R-CDFR uses the TRADES objective (Zhang et al., 2019) as its base adversarial-training loss and adds the three regularizers in Eq. (12). The main backbone is WRN-34-10 (Zagoruyko & Komodakis, 2016); additional ResNet-18 (He et al., 2016) and DeiT-S (Touvron et al., 2021) results are reported in Appendix F. Training details, hyperparameters, evaluator settings, and compute measurements are provided in Appendix B.

We compare R-CDFR with **classical AT methods** (PGD-AT (Madry et al., 2018), TRADES (Zhang et al., 2019), MART (Wang et al., 2020), and AWP (Wu et al., 2020)), **stability-based regularizers** (Parseval (Cisse et al., 2017), Jacobian regularization (Jakubovitz & Giryes, 2018), and Lipschitz-margin training (Tsuzuku et al., 2018)), and **geometry-aware methods** (TangentAT (Yu et al., 2019), TART (Yi et al., 2024), Consistency AT (Tack et al., 2022), and MAT (Zhang et al., 2018)). We also include an autoencoder-projection (AE-Proj) baseline inspired by manifold projection and purification (Samangouei et al., 2018; Stutz et al., 2019). All baselines use the same backbone, preprocessing, optimizer family, training budget, and evaluation protocol. Unless stated otherwise, all results are reported as mean $\pm$ standard deviation over five runs.

6.2 RQ-A: STANDARD ROBUSTNESS AND COMPLEMENTARY EVALUATIONS

Table 1 reports matched WRN-34-10 results on CIFAR-10 and CIFAR-100. R-CDFR achieves the highest AutoAttack accuracy on both datasets, with the same ordering on TinyImageNet (Appendix

Group	Method	CIFAR-10			CIFAR-100		
		Clean↑	PGD-20↑	AA↑	Clean↑	PGD-20↑	AA↑
No AT	Standard ERM	**93.90±0.21**	0.31±0.09	0.08±0.03	**76.42±0.35**	0.16±0.05	0.04±0.02
AT	PGD-AT	62.86±0.47	56.17±0.71	47.34±0.83	58.79±0.52	32.36±0.94	28.94±1.02
	TRADES	78.41±0.39	33.67±0.88	31.25±0.91	68.86±0.44	25.08±0.77	23.12±0.81
	MART	74.06±0.43	57.12±0.69	51.27±0.76	64.31±0.50	34.38±0.89	31.19±0.93
	AWP	66.99±0.51	60.63±0.74	55.31±0.82	61.17±0.46	36.24±0.85	33.41±0.88
Stability	Parseval	93.15±0.18	13.28±0.57	11.06±0.52	75.68±0.31	6.69±0.44	5.54±0.41
	Jacobian Reg.	92.91±0.19	15.42±0.61	12.87±0.58	75.12±0.33	7.48±0.48	6.21±0.45
	Lipschitz-Margin	92.76±0.22	17.13±0.64	14.62±0.60	74.89±0.36	8.26±0.51	6.97±0.49
Geometry	TangentAT	77.12±0.41	58.24±0.66	53.49±0.74	65.54±0.48	35.15±0.82	32.28±0.86
	TART	78.49±0.38	60.07±0.71	55.38±0.79	66.61±0.45	36.09±0.88	33.20±0.91
	Consistency AT	78.16±0.40	59.31±0.69	54.67±0.76	66.18±0.47	35.62±0.84	32.71±0.89
	MAT	76.28±0.43	57.15±0.72	51.93±0.81	64.27±0.51	33.69±0.90	30.47±0.95
	AE-Proj	86.57±0.27	45.26±0.83	40.18±0.86	70.46±0.39	23.27±0.71	20.59±0.75
Ours	**R-CDFR**	79.97±0.36	**61.63±0.65**	**57.43±0.71**	67.09±0.44	**39.15±0.74**	**36.98±0.78**

Table 1: Matched WRN-34-10 results on CIFAR-10 and CIFAR-100. Robust accuracy is evaluated under ℓ_∞, $\epsilon = 8/255$.

Method	CIFAR-10			CIFAR-100		
	$S_P\downarrow$	$S_R\downarrow$	$G_P\downarrow$	$S_P\downarrow$	$S_R\downarrow$	$G_P\downarrow$
PGD-AT	1.46±0.05	2.08±0.07	1.21±0.04	2.32±0.08	3.43±0.11	1.96±0.07
TRADES	1.72±0.06	2.35±0.08	1.38±0.05	2.74±0.09	3.82±0.12	2.18±0.08
MART	1.39±0.05	1.96±0.06	1.14±0.04	2.18±0.07	3.15±0.10	1.84±0.06
AWP	1.22±0.04	1.78±0.05	1.02±0.03	1.96±0.07	2.87±0.09	1.71±0.06
Parseval	2.86±0.10	4.12±0.13	2.41±0.08	4.21±0.15	5.98±0.18	3.52±0.11
Jacobian Reg.	2.41±0.09	3.78±0.12	2.05±0.07	3.74±0.13	5.41±0.17	3.07±0.10
Lipschitz-Margin	2.18±0.08	3.54±0.11	1.92±0.07	3.42±0.12	5.12±0.16	2.86±0.09
TangentAT	1.34±0.05	1.42±0.05	0.94±0.03	2.04±0.07	2.31±0.08	1.54±0.05
TART	1.29±0.04	1.36±0.05	0.89±0.03	1.98±0.06	2.24±0.08	1.49±0.05
Consistency AT	1.31±0.04	1.40±0.05	0.86±0.03	2.01±0.07	2.29±0.08	1.43±0.05
MAT	1.38±0.05	1.55±0.06	0.98±0.04	2.12±0.07	2.46±0.09	1.61±0.06
AE-Proj	1.67±0.06	1.92±0.07	1.23±0.05	2.48±0.08	2.95±0.10	1.98±0.07
R-CDFR	**0.91±0.03**	**1.08±0.04**	**0.62±0.02**	**1.47±0.05**	**1.71±0.07**	**1.06±0.04**

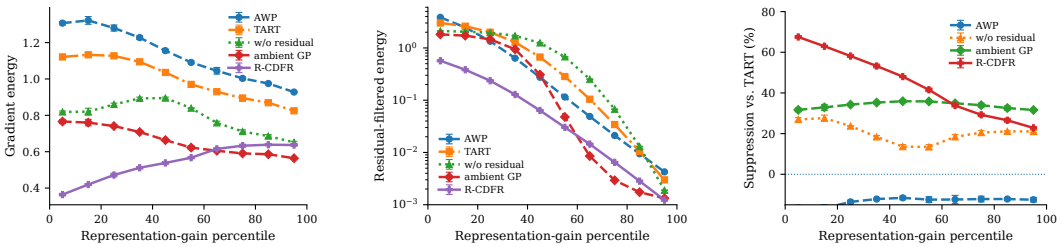
Table 2: Model-relative diagnostics on CIFAR-10 and CIFAR-100. S_P is pass-filtered loss sensitivity, S_R is residual-filtered loss sensitivity, and G_P is predictive KL divergence under a small pass-filtered perturbation. TinyImageNet results are reported in Appendix J.2.

Table 6). The gains persist when R-CDFR is added to a stronger TRADES-AWP recipe and when evaluated on DeiT-S and ResNet-18 (Appendix Tables 10, 11, and 12). Training trajectories are reported in Appendix D.

We further evaluate robustness beyond the main ℓ_∞ setting. Appendix Table 14 shows that R-CDFR retains the strongest robust accuracy under both PGD-ℓ_2 and AutoAttack-ℓ_2 across all three datasets. The fixed-reference pass/residual stress test also preserves the ordering under a shared external encoder (Appendix Table 15). Finally, Appendix Table 16 reports corruption accuracy and mean corruption error (mCE), where lower mCE indicates stronger corruption robustness. R-CDFR achieves the lowest mCE and the highest corruption accuracy among the adversarial-training and geometry-aware baselines. These results show that the improvement persists across threat models, training recipes, architectures, external filter-aligned attacks, and natural distribution shifts.

6.3 RQ-B: Filtered Loss Coefficients and Finite-Step Consistency

To examine whether the proposed objective reduces the quantities identified by the local analysis, we measure the pass coefficient $S_P = \|P_\lambda g\|_{A_\lambda^{-1}}$, the residual coefficient $S_R = \|R_\lambda g\|_2$, and the finite-step divergence $G_P = \mathrm{KL}(p_\theta(x)\|p_\theta(x+\delta_P))$. The first two are the local first-order coefficients in Proposition 5.1, while G_P measures the complementary finite-step behavior targeted



(a) Gradient energy. (b) Residual-filtered energy. (c) Suppression relative to TART.

Figure 2: Gradient allocation across the local representation-gain spectrum. For each sample, the nonzero singular directions of the representation Jacobian are divided into ten equal-count bins according to their within-sample gain percentile. (a) Mean input-gradient energy a_i^2, where $a_i = v_i^\top g$. (b) Mean residual-filtered energy $\big(\lambda/(\sigma_i^2+\lambda)\big)^2 a_i^2$ on a logarithmic scale. (c) Suppression of the unfiltered gradient energy relative to TART. Panels (a) and (b) are normalized by the corresponding mean TART energy across bins.

by $\mathcal{L}_{\text{cons}}$. Table 2 shows that R-CDFR obtains the lowest value of all three quantities on CIFAR-10 and CIFAR-100, with the same ordering on TinyImageNet (Appendix Table 19).

The fixed- and cross-representation controls in Appendix I.1 preserve this ordering under a stationary or external encoder. The consistency curves in Appendix J.1 further show a widening KL advantage as the pass-filtered step grows. Together, these results support the intended alignment between the three regularizers and their corresponding local and finite-step diagnostics, while the robustness outcomes are evaluated separately in RQ-A.

6.4 RQ-C: Gain Selectivity and Residual-Direction Behavior

To examine whether the reduction in filtered sensitivity reflects a gain-selective change in the input gradient, rather than uniform gradient shrinkage, we analyze gradient energy across the local representation-gain spectrum. For each test sample, we sort the nonzero singular directions $\{v_i\}_{i=1}^r$ by their directional gains σ_i and partition them into ten equal-count percentile bins $\{\mathcal{B}_b\}_{b=1}^{10}$. With $a_i = v_i^\top g$, we compute the mean gradient energy and residual-filtered energy in bin b as

$$E_b^{\text{grad}} = \mathbb{E}_x\left[\frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b} a_i^2\right], \qquad E_b^{\text{res}} = \mathbb{E}_x\left[\frac{1}{|\mathcal{B}_b|}\sum_{i\in\mathcal{B}_b}\left(\frac{\lambda}{\sigma_i^2+\lambda}\right)^2 a_i^2\right]. \tag{15}$$

We additionally report the bin-wise suppression relative to TART,

$$Q_b(m) = 100\left(1 - \frac{E_b^{\text{grad}}(m)}{E_b^{\text{grad}}(\text{TART})}\right), \tag{16}$$

where m denotes the evaluated method. Figure 2 shows a distinctly gain-selective response. R-CDFR suppresses substantially more unfiltered gradient energy in the low-gain bins than in the high-gain bins, whereas the intensity-matched ambient gradient penalty produces a much flatter profile. Removing $\mathcal{L}_{\text{res}}$ weakens this low-to-high gain contrast. The spectrum-resolved result therefore supports the central mechanism claim: the improvement is associated with continuous low-gain emphasis rather than uniform gradient shrinkage.

Residual energy and decision distance. For each sample, define the residual energy $D_R(x) = \|R_\lambda \nabla_x \ell(x,y)\|_2^2$. We also normalize the residual-filtered gradient and line-search for the first label change along that direction.The former measures local low-gain loss sensitivity; the latter measures a one-dimensional decision-changing distance under the model-induced filter.

Figure 3 shows a right shift for R-CDFR relative to PGD-AT, TRADES, and TART, so a larger step is required to change the prediction along the selected residual-filtered direction. Together with the lower residual coefficient S_R reported in Table 2, this right shift connects reduced local residual sensitivity with a larger one-dimensional decision-changing displacement under the same learned filter. Matched controls distinguish this behavior from simpler alternatives. Appendix Table 8 shows that R-CDFR outperforms intensity-matched ambient gradient regularization, while

Dataset	Variant	Clean↑	AA↑	$S_P \downarrow$	$S_R \downarrow$	$G_P \downarrow$	$D_R \downarrow$
CIFAR-10	**R-CDFR**	79.97±0.36	**57.43±0.71**	**0.91±0.03**	**1.08±0.04**	**0.62±0.02**	**0.020±0.003**
	w/o $\mathcal{L}_{\text{spec}}$	79.85±0.35	52.14±0.73	1.27±0.05	1.22±0.05	0.76±0.03	0.030±0.004
	w/o $\mathcal{L}_{\text{res}}$	**80.12±0.33**	49.38±0.68	1.08±0.04	1.58±0.06	0.71±0.03	0.070±0.006
	w/o $\mathcal{L}_{\text{cons}}$	79.91±0.34	53.47±0.70	0.96±0.04	1.13±0.04	0.94±0.04	0.020±0.003
CIFAR-100	**R-CDFR**	67.09±0.44	**36.98±0.78**	**1.47±0.05**	**1.71±0.07**	**1.06±0.04**	**0.030±0.004**
	w/o $\mathcal{L}_{\text{spec}}$	66.92±0.42	32.41±0.80	2.06±0.08	1.92±0.08	1.29±0.05	0.040±0.005
	w/o $\mathcal{L}_{\text{res}}$	**67.18±0.40**	29.26±0.74	1.74±0.06	2.42±0.09	1.20±0.05	0.110±0.010
	w/o $\mathcal{L}_{\text{cons}}$	67.02±0.41	33.58±0.77	1.55±0.06	1.79±0.07	1.63±0.06	0.040±0.005
TinyImageNet	**R-CDFR**	54.48±0.43	**30.82±0.90**	**1.86±0.06**	**2.21±0.08**	**1.28±0.05**	**0.050±0.006**
	w/o $\mathcal{L}_{\text{spec}}$	54.31±0.45	26.73±0.93	2.59±0.09	2.48±0.09	1.56±0.06	0.070±0.008
	w/o $\mathcal{L}_{\text{res}}$	**54.66±0.44**	23.81±0.88	2.18±0.08	3.14±0.11	1.47±0.06	0.160±0.014
	w/o $\mathcal{L}_{\text{cons}}$	54.39±0.46	27.64±0.91	1.97±0.07	2.34±0.09	2.01±0.08	0.060±0.007

Table 3: Component ablation on CIFAR-10, CIFAR-100, and TinyImageNet. S_P and S_R are the pass- and residual-filtered first-order coefficients, G_P is the pass-filtered predictive KL divergence, and $D_R = \|R_\lambda g\|_2^2$ is the residual-filtered gradient energy.

Appendix Table 9 shows that it also improves over hard spectral partitioning. Appendix Table 13 further demonstrates the strongest directional robustness under known geometry, supporting the role of low-gain residual suppression.

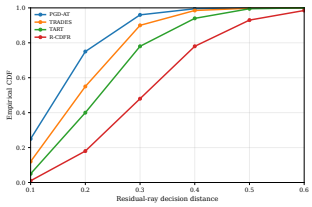


Figure 3: CDF of residual-ray decision distances on CIFAR-10. Right-shifted curves indicate larger label changing steps along normalized residual filtered gradients.

6.5 RQ-D: Component Ablation

We evaluate each regularizer using AutoAttack accuracy and its associated mechanism-specific diagnostic. Table 3 reports the complete component ablation on all three datasets. Removing $\mathcal{L}_{\text{spec}}$ produces the largest increase in S_P on all three datasets, consistent with its role in spectrum-wide slope control. Removing $\mathcal{L}_{\text{res}}$ produces the largest increases in S_R and D_R and causes the largest decrease in AutoAttack accuracy, showing that concentrated low-gain suppression is central to the robustness gain. Removing $\mathcal{L}_{\text{cons}}$ primarily increases G_P, while leaving D_R relatively close to the full objective, which supports its distinct role in finite-step predictive stability. All three removals reduce robust accuracy, whereas the clean accuracies remain within 0.2 points of the full model. The term-specific changes therefore support the intended correspondence between $\mathcal{L}_{\text{spec}}$, $\mathcal{L}_{\text{res}}$, and $\mathcal{L}_{\text{cons}}$ and their respective spectrum-wide, residual, and finite-step effects. Weight and damping analyses are reported in Appendices E.1–E.4.

7 Conclusion

R-CDFR uses local representation gain as a continuous signal for allocating adversarial regularization. Complementary damped filters define pass and residual responses without selecting a rank or imposing a binary subspace split. The resulting objective combines broad inverse-metric slope control, sharper low-gain residual suppression, and finite-step predictive consistency. The local analysis identifies the pass and residual first-order coefficients induced by these filters, while the experiments show that the optimized quantities, the spectrum-resolved gradient allocation, and the robustness outcome move together. The same pattern appears under matched standard protocols, a stronger TRADES-AWP recipe, convolutional and transformer backbones, and analytic tasks with known directional geometry.

Reproducibility Statement

We provide the training and evaluation protocols, hyperparameters, attack settings, implementation details for the damped operators, and computational configuration in the appendix. The final artifact will release the code, configuration files, random seeds, and scripts needed to reproduce the main experiments, ablations, diagnostics, and reported tables and figures.

Ethics Statement

This work studies adversarial robustness using standard public image benchmarks and synthetic tasks with known geometry. It involves no human subjects, sensitive personal data, or deployment-specific decision making. R-CDFR is intended to improve the robustness of learned models against adversarial perturbations. The reported empirical gains should be interpreted within the evaluated threat models and do not constitute a guarantee of robustness or safety in deployment.

AI Use Statement

Generative AI tools assisted with language editing, code drafting, and consistency checks. The authors remain responsible for all claims, citations, experimental results, and released artifacts, and will verify every AI-assisted component before submission.

References

Sajjad Amini, Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *arXiv preprint arXiv:2406.05927*, 2024.

Yatong Bai, Mo Zhou, Vishal M Patel, and Somayeh Sojoudi. Mixednuts: Training-free accuracy-robustness balance via nonlinearly mixed classifiers. *arXiv preprint arXiv:2402.02263*, 2024.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7 (85):2399–2434, 2006.

Alvin Chan, Yi Tay Chen, and Chee Seng Chan. Jacobian adversarially regularized networks for robustness. In *International Conference on Learning Representations (ICLR)*, 2020.

Erh-Chung Chen, Pin-Yu Chen, I-Hsin Chung, and Che-Rung Lee. Data-driven lipschitz continuity: A cost-effective approach to improve adversarial robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Jiahao Chen, Zhou Feng, Rui Zeng, Yuwen Pu, Chunyi Zhou, Yi Jiang, Yuyou Gan, Jinbao Li, and Shouling Ji. Enhancing adversarial transferability with adversarial weight tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

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, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, Aleksander Madry, and Matthias Hein. Robustbench: A standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Elvis Dohmatob. Limitations of adversarial robustness: Strong no free lunch theorem. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chris Finlay, Jeff Calder, Bilal Abbasi, and Adam Oberman. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*, 2018a.

Chris Finlay, Adam M Oberman, and Bilal Abbasi. Improved robustness to adversarial examples using lipschitz regularization of the loss. In *arXiv preprint arXiv:1810.00953*, 2018b.

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.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Yueru Li, Shuyu Cheng, Hang Su, and Jun Zhu. Defense against adversarial attacks via controlling gradient leaking on embedded manifolds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Zhiting Li, Shibai Yin, Tai-Xiang Jiang, Yexun Hu, Jia-Mian Wu, Guowei Yang, and Guisong Liu. Enhancing the adversarial robustness via manifold projection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

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 (ICLR)*, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

Ramchandran Muthukumar and Jeremias Sulam. Adversarial robustness of sparse local lipschitz predictors. *SIAM Journal on Mathematics of Data Science*, 5(4):920–948, 2023.

Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations (ICLR)*, 2018.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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 (AAAI)*, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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 (ICLR)*, 2020.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *International Conference on Machine Learning (ICML)*, 2023.

Eric Wong, Leslie Rice, and Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tao Wu, Tie Luo, and Donald C Wunsch II. Lrs: Enhancing adversarial transferability through lipschitz regularized surrogate. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Jiancong Xiao, Liusha Yang, Yanbo Fan, Jue Wang, and Zhi-Quan Luo. Understanding adversarial robustness against on-manifold adversarial examples. *Pattern Recognition*, 159:111071, 2025.

Bongsoo Yi, Rongjie Lai, and Yao Li. Tart: Boosting clean accuracy through tangent direction guided adversarial training. *arXiv preprint arXiv:2408.14728*, 2024.

Bing Yu, Jingfeng Wu, Jinwen Ma, and Zhanxing Zhu. Tangent-normal adversarial regularization for semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Lilin Zhang, Chengpei Wu, and Ning Yang. Weakly supervised contrastive adversarial training for learning robust features from semi-supervised data. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

Shufei Zhang, Kaizhu Huang, Jianke Zhu, and Yang Liu. Manifold adversarial learning. In *arXiv preprint arXiv:1807.05832*, 2018.

Wenjia Zhang, Yikai Zhang, Xiaoling Hu, Mayank Goswami, Chao Chen, and Dimitris N Metaxas. A manifold view of adversarial risk. In *International Conference on Machine Learning (ICML)*, 2022b.

Wenjia Zhang, Yikai Zhang, Xiaoling Hu, Yi Yao, Mayank Goswami, Chao Chen, and Dimitris Metaxas. Manifold-driven decomposition for adversarial robustness. *Frontiers in Computer Science*, 5:1274695, 2024.

Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of lipschitz calculus: A survey. *ACM Computing Surveys*, 57(6):1–41, 2025.

Algorithm 1 R-CDFR training for one mini-batch

Require: Mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$; model $f_\theta = c_\phi \circ h_\psi$; damping λ; weights α, β, γ; pass radius ε_P; numerical constant η

Compute the base mini-batch objective $\mathcal{L}_{\text{base}}$
for $i = 1, \dots, B$ **do**
    Compute $\ell_i = \ell(x_i, y_i)$ and $g_i = \nabla_{x_i} \ell_i$
    Let $J_i = J_h(x_i)$ and implement products with J_i and $J_i^\top$ using automatic differentiation
    Solve $(J_i^\top J_i + \lambda I_d) w_i = g_i$
    Set $s_i = \sqrt{g_i^\top w_i}$ and $d_i = \|\lambda w_i\|_2^2$
    Sample $u_i \sim \mathcal{N}(0, I_d)$
    Solve $(J_i J_i^\top + \lambda I_m) t_i = J_i u_i$ and set $\widetilde{u}_i = J_i^\top t_i$
    Form $\delta_{P,i} = \varepsilon_P \widetilde{u}_i / (\|\widetilde{u}_i\|_2 + \eta)$
    Set $c_i = \mathrm{KL}\big(p_\theta(x_i) \,\|\, p_\theta(x_i + \delta_{P,i})\big)$
end for
Set $\mathcal{L}_{\text{spec}} = B^{-1} \sum_{i=1}^{B} s_i$, $\mathcal{L}_{\text{res}} = B^{-1} \sum_{i=1}^{B} d_i$, and $\mathcal{L}_{\text{cons}} = B^{-1} \sum_{i=1}^{B} c_i$
Update θ using Eq. (12)

A Proof of Proposition 5.1

Fix $\delta = P_\lambda u_P + R_\lambda u_R \in \Delta_\lambda(x; \tau_P, \varepsilon_R)$. The H-Lipschitz continuity of the input gradient gives

$$\ell(x + \delta, y) \le \ell(x, y) + g^\top \delta + \frac{H}{2} \|\delta\|_2^2 . \tag{17}$$

Because P_λ and R_λ are symmetric,

$$g^\top P_\lambda u_P = (P_\lambda g)^\top u_P \le \|P_\lambda g\|_{A_\lambda^{-1}} \|u_P\|_{A_\lambda} ,$$
$$g^\top R_\lambda u_R = (R_\lambda g)^\top u_R \le \|R_\lambda g\|_2 \|u_R\|_2 .$$

Using the constraints in Eq. (13), the linear term satisfies

$$g^\top \delta \le \tau_P \|P_\lambda g\|_{A_\lambda^{-1}} + \varepsilon_R \|R_\lambda g\|_2 = \tau_P S_P(x) + \varepsilon_R S_R(x).$$

Moreover, $A_\lambda \succeq \lambda I_d$ gives $\|u_P\|_2 \le \tau_P / \sqrt{\lambda}$. Since both filters are non-expansive in the Euclidean norm,

$$\|\delta\|_2 \le \|P_\lambda u_P\|_2 + \|R_\lambda u_R\|_2 \le \frac{\tau_P}{\sqrt{\lambda}} + \varepsilon_R.$$

Substituting these bounds into Eq. (17) and taking the supremum over $\delta \in \Delta_\lambda(x; \tau_P, \varepsilon_R)$ proves Eq. (14).

B Implementation and Training Details

B.1 R-CDFR Training Algorithm

B.2 Default hyperparameters

Dataset	α	β	γ	λ	ε_P
CIFAR-10	1.0	2.0	0.5	10^{-3}	0.25
CIFAR-100	1.5	3.0	0.5	10^{-3}	0.20
TinyImageNet	2.0	4.0	1.0	10^{-3}	0.15

Table 4: Default weights used in the matched training protocol. α, β, and γ weight the spectral-slope, residual-filter, and finite-step consistency terms.

The default convolutional experiments use WRN-34-10, 50 epochs, and the same preprocessing, optimizer family, batch size, and validation protocol for all matched baselines. CIFAR-10/100 results

use five independent training seeds, while TinyImageNet uses three. The strong-recipe experiment uses 200 epochs and three seeds. The primary objective forms its filters from the same internal representation h_ψ used by the classifier. Fixed-reference settings use separately specified encoders and are reported as external controls.

B.3 Linear solves and differentiation

For $P_\lambda v$, solve $(JJ^\top + \lambda I)s = Jv$ by conjugate gradients and return $J^\top s$. For $\|g\|_{A_\lambda^{-1}}$, solve $A_\lambda w = g$ using products $v \mapsto J^\top(Jv) + \lambda v$. The default is 10 CG iterations; relative residuals are typically below 10^{-3}. JVP/VJP operations avoid explicit Jacobians. Training gradients through $g = \nabla_x \ell$ and through the solves use autograd with create-graph support; this is higher-order differentiation even though no Hessian is materialized.

Method	Time/epoch	Peak memory/GPU	Relative time
TRADES	1.92 min	10.8 GB	1.00×
AWP	2.07 min	11.6 GB	1.08×
R-CDFR	2.29 min	12.4 GB	1.19×
TART	2.78 min	13.7 GB	1.45×

Table 5: Matched WRN-34-10 compute on two A100-40GB GPUs.

B.4 Evaluator specification

All ℓ_∞ attacks receive raw $[0, 1]$ tensors. Channel normalization is part of the attacked model wrapper, so the perturbation constraint is applied in raw pixel coordinates. AutoAttack uses the official default ensemble on the complete test set. For ERM, PGD-20 and AA are 0.31 ± 0.09 and 0.08 ± 0.03 on CIFAR-10. The corresponding values are 0.16 ± 0.05 and 0.04 ± 0.02 on CIFAR-100, and 0.23 ± 0.07 and 0.06 ± 0.02 on TinyImageNet.

C Additional Standard Robustness Results

Table 6 reports the TinyImageNet results corresponding to the CIFAR-10/100 comparison in Table 1.

Group	Method	Clean	PGD-20	AA
No AT	Standard ERM	$\mathbf{67.84\pm0.42}$	0.23 ± 0.07	0.06 ± 0.02
AT	PGD-AT	53.97 ± 0.53	28.46 ± 0.97	25.12 ± 1.05
	TRADES	56.21 ± 0.48	19.36 ± 0.82	17.92 ± 0.86
	MART	55.08 ± 0.51	30.28 ± 0.94	27.14 ± 0.99
	AWP	54.62 ± 0.47	31.96 ± 0.91	29.37 ± 0.95
Stability	Parseval	66.95 ± 0.36	7.86 ± 0.54	6.54 ± 0.50
	Jacobian Reg.	66.12 ± 0.38	9.41 ± 0.59	7.83 ± 0.56
	Lipschitz-Margin	65.87 ± 0.41	11.02 ± 0.63	9.24 ± 0.60
Geometry	TangentAT	52.71 ± 0.49	30.11 ± 0.88	27.65 ± 0.92
	TART	55.12 ± 0.47	31.48 ± 0.91	28.94 ± 0.96
	Consistency AT	53.21 ± 0.46	30.92 ± 0.89	28.41 ± 0.94
	MAT	51.84 ± 0.52	29.17 ± 0.95	26.63 ± 0.99
	AE-Proj	59.12 ± 0.44	18.73 ± 0.78	16.52 ± 0.82
Ours	**R-CDFR**	54.48 ± 0.43	$\mathbf{33.06\pm0.86}$	$\mathbf{30.82\pm0.90}$

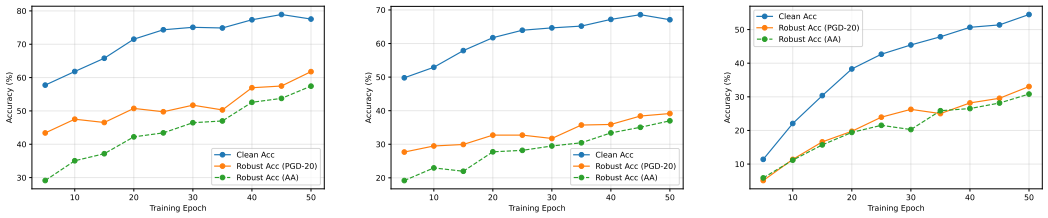
Table 6: Matched WRN-34-10 results on TinyImageNet over three seeds under ℓ_∞, $\epsilon = 8/255$.

D Training Dynamics

Figure 4 reports clean, PGD-20, and AutoAttack accuracy throughout training. CIFAR-10 reaches its robust plateau earlier, whereas CIFAR-100 and TinyImageNet improve more gradually. On all three datasets, PGD-20 and AutoAttack rise through the final checkpoints, with AutoAttack remaining consistently below PGD-20.

m_β	CIFAR-10 Clean	CIFAR-10 AA	CIFAR-100 Clean	CIFAR-100 AA
0	80.12±0.33	49.38±0.68	67.18±0.40	29.26±0.74
0.25	80.09±0.34	55.22±0.72	67.16±0.41	32.08±0.77
0.50	80.04±0.35	56.55±0.70	67.13±0.42	35.48±0.76
1.00	79.97±0.36	**57.43±0.71**	67.09±0.44	**36.98±0.78**
2.00	79.45±0.41	57.02±0.75	66.64±0.47	35.87±0.80

Table 7: One-factor sweep of $m_\beta = \beta/\beta_0$, with $\beta_0 = 2$ on CIFAR-10 and 3 on CIFAR-100. Values are mean ± standard deviation over three independent runs.



(a) CIFAR-10 (b) CIFAR-100 (c) TinyImageNet

Figure 4: Clean, PGD-20, and AutoAttack accuracy over training. Robust accuracy is evaluated every five epochs. CIFAR-10 converges earlier, while CIFAR-100 and TinyImageNet exhibit slower but sustained robust improvement.

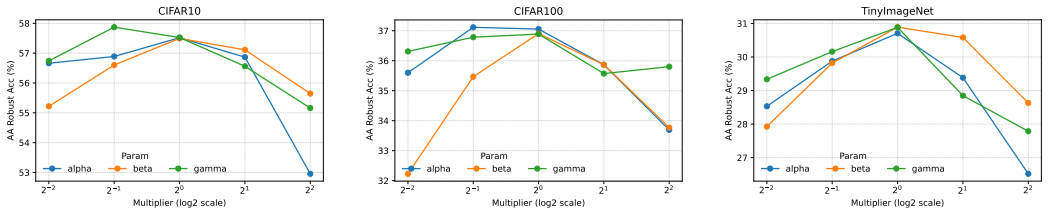
Near convergence, the gap between PGD-20 and AutoAttack is stable and both curves retain the same direction of improvement. TinyImageNet exhibits larger fluctuations, consistent with its smaller per-class sample count and more difficult optimization regime. The trajectories therefore support the final-checkpoint comparison rather than indicating an isolated late-epoch gain.

E Ablations and Hyperparameter Sensitivity

E.1 Residual-weight sweep

Figure 5 reports one-factor AutoAttack sweeps for the three objective weights, and Table 7 adds clean accuracy for the residual-weight slice. The multiplier is applied relative to the dataset-specific default in Table 4.

All three datasets exhibit a broad high-accuracy region around the default multiplier. AutoAttack decreases when α is too small and again at the largest tested value, matching the need for spectrum-wide control without excessive regularization. The β curve is especially influential on CIFAR-100, consistent with the larger residual-ablation gap in Table 3. The non-monotone γ curve similarly places the useful finite-step consistency range around the default. Table 7 makes the clean–robust operating point explicit: moving from $m_\beta = 0$ to the default raises AutoAttack by 8.05 points on CIFAR-10 and 7.72 points on CIFAR-100, while clean accuracy changes by -0.15 and -0.09 points. Doubling the default weight lowers both clean and robust accuracy, indicating that the observed gain is associated with a finite range of residual emphasis rather than monotonic gradient suppression.



(a) CIFAR-10 (b) CIFAR-100 (c) TinyImageNet

Figure 5: AutoAttack sensitivity to the spectral-slope weight α, residual-filter weight β, and consistency weight γ. One weight is varied while the other two remain at their defaults; the horizontal axis is a base-2 multiplier.

E.2 Matched Ambient-Gradient Control

To distinguish gain-conditioned residual regularization from uniform gradient shrinkage, we replace $\mathcal{L}_{\text{res}}$ with the ambient gradient penalty

$$\mathcal{L}_{\text{GP}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|g\|_2^2\right]. \tag{18}$$

The coefficient of $\mathcal{L}_{\text{GP}}$ is calibrated on a fixed set of training mini-batches so that its initial weighted magnitude matches that of $\beta\mathcal{L}_{\text{res}}$:

$$\beta_{\text{GP}} = \beta\frac{\mathbb{E}\left[\|R_\lambda g\|_2^2\right]}{\mathbb{E}\left[\|g\|_2^2\right]}. \tag{19}$$

All other components, including the TRADES base objective, $\mathcal{L}_{\text{spec}}$, $\mathcal{L}_{\text{cons}}$, architecture, training schedule, and checkpoint selection, are held fixed.

Table 8 compares the damped residual penalty with an intensity-matched ambient gradient penalty under otherwise identical training settings. The ambient-GP variant improves AutoAttack accuracy over the residual ablation, confirming that aggregate gradient control is beneficial. However, R-CDFR further improves AutoAttack by 2.52 points on CIFAR-10 and 2.70 points on CIFAR-100 relative to ambient GP, while maintaining comparable clean accuracy. The difference is more pronounced in the residual-specific diagnostics: R-CDFR achieves substantially lower S_R and D_R, together with lower S_P and G_P. These results indicate that the observed improvement is not explained by uniform gradient shrinkage alone and support the role of representation-conditioned residual weighting.

E.3 Damped versus Hard Spectral Splits

To isolate the effect of continuous spectral weighting, we compare R-CDFR with two hard partitions of the same representation-gain spectrum. Given the singular directions $\{v_i\}_{i=1}^r$ and gains $\{\sigma_i\}_{i=1}^r$, the **hard-threshold filter** is

$$P_{\text{thr}} = \sum_{i=1}^{r}\mathbb{I}\left[\sigma_i^2 \geq \lambda\right] v_i v_i^\top, \qquad R_{\text{thr}} = I_d - P_{\text{thr}}, \tag{20}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

The **rank-matched top-k filter** uses

$$k(x) = \text{round}\left(\text{tr}(P_\lambda(x))\right), \qquad P_k = \sum_{i=1}^{k(x)} v_i v_i^\top, \qquad R_k = I_d - P_k. \tag{21}$$

Thus, P_k matches the effective pass dimension of the damped filter at each input while replacing its continuous responses with binary weights.

Dataset	Variant	Clean↑	AA↑	$S_P \downarrow$	$S_R \downarrow$	$G_P \downarrow$	$D_R \downarrow$
CIFAR-10	w/o residual	**80.12±0.33**	49.38±0.68	1.08±0.04	1.58±0.06	0.71±0.03	0.070±0.006
	Ambient GP	80.04±0.31	54.91±0.61	0.99±0.04	1.30±0.05	0.66±0.03	0.043±0.004
	R-CDFR	79.97±0.36	**57.43±0.71**	**0.91±0.03**	**1.08±0.04**	**0.62±0.02**	**0.020±0.003**
CIFAR-100	w/o residual	**67.18±0.40**	29.26±0.74	1.74±0.06	2.42±0.09	1.20±0.05	0.110±0.010
	Ambient GP	67.14±0.38	34.28±0.69	1.60±0.05	2.01±0.08	1.12±0.04	0.064±0.006
	R-CDFR	67.09±0.44	**36.98±0.78**	**1.47±0.05**	**1.71±0.07**	**1.06±0.04**	**0.030±0.004**

Table 8: Matched natural-image control for ambient gradient regularization. The ambient-GP variant replaces $\mathcal{L}_{\mathrm{res}}$ with the intensity-matched penalty $\mathcal{L}_{\mathrm{GP}} = \mathbb{E}\,\|g\|_2^2$, while all other objective terms and training settings remain unchanged. S_P and S_R denote pass- and residual-filtered loss sensitivity, G_P denotes predictive KL divergence under a fixed pass-filtered step, and $D_R = \|R_\lambda g\|_2^2$ denotes residual-filtered gradient energy. Values are mean ± standard deviation over three seeds.

Dataset	Variant	Clean↑	AA↑	$S_P \downarrow$	$S_R \downarrow$	$G_P \downarrow$	$D_R \downarrow$	$\Delta Q \uparrow$
CIFAR-10	Hard threshold	**80.01±0.34**	55.86±0.74	0.97±0.04	1.21±0.05	0.68±0.03	0.033±0.004	27.6±2.0
	Rank-matched top-k	79.95±0.35	56.34±0.72	0.95±0.04	1.16±0.05	0.66±0.03	0.028±0.004	31.9±1.9
	R-CDFR	79.97±0.36	**57.43±0.71**	**0.91±0.03**	**1.08±0.04**	**0.62±0.02**	**0.020±0.003**	**39.8±1.7**
CIFAR-100	Hard threshold	**67.15±0.41**	35.06±0.81	1.58±0.06	1.95±0.08	1.15±0.05	0.051±0.006	26.1±2.3
	Rank-matched top-k	67.11±0.42	35.61±0.80	1.54±0.05	1.86±0.08	1.12±0.04	0.044±0.005	31.4±2.1
	R-CDFR	67.09±0.44	**36.98±0.78**	**1.47±0.05**	**1.71±0.07**	**1.06±0.04**	**0.030±0.004**	**42.3±2.0**

Table 9: Comparison of continuous damped filtering with hard spectral partitions. The hard-threshold control binarizes the spectral response at $\sigma_i^2 = \lambda$, while the rank-matched top-k control matches the effective pass dimension $\mathrm{tr}(P_\lambda)$. Residual-penalty magnitudes are matched at initialization, and all other objective terms and training settings remain fixed. The diagnostics are evaluated using the same damped operators for all variants. ΔQ is the difference between average suppression in the lowest and highest three gain bins. Values are mean ± standard deviation over three seeds.

For each hard-filter control, we replace P_λ and R_λ in $\mathcal{L}_{\mathrm{cons}}$ and $\mathcal{L}_{\mathrm{res}}$, respectively, while retaining the TRADES base objective and $\mathcal{L}_{\mathrm{spec}}$. To remove differences caused only by the scale of the residual penalty, the hard-filter coefficient is calibrated on fixed training mini-batches as

$$\beta_H = \beta \frac{\mathbb{E}\left[\|R_\lambda g\|_2^2\right]}{\mathbb{E}\left[\|R_H g\|_2^2\right]}, \qquad R_H \in \{R_{\mathrm{thr}}, R_k\}. \tag{22}$$

All other training settings are unchanged.

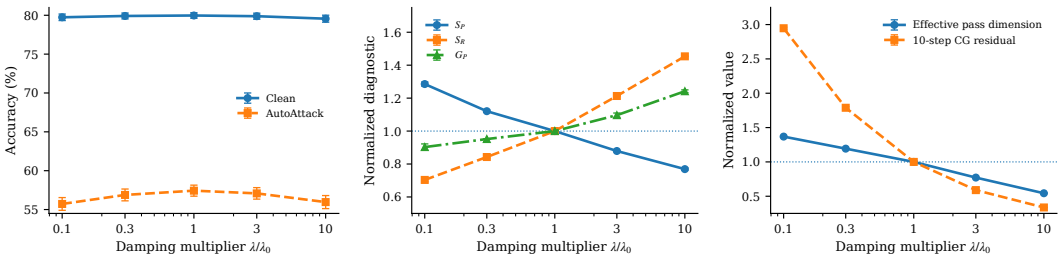
For comparable evaluation, S_P, S_R, G_P, and D_R are computed using the damped operators in Eq. (4) for every trained variant, irrespective of the filter used during training. We also summarize spectral selectivity by

$$\Delta Q := Q_{\mathrm{low}} - Q_{\mathrm{high}}, \qquad Q_{\mathrm{low}} := \frac{1}{3}\sum_{b=1}^{3} Q_b, \qquad Q_{\mathrm{high}} := \frac{1}{3}\sum_{b=8}^{10} Q_b, \tag{23}$$

where Q_b is the bin-wise suppression defined in Eq. (16). A larger ΔQ indicates that suppression is more concentrated toward the low-gain portion of the spectrum.

Table 9 compares continuous damping with two binary spectral partitions under matched training conditions. Both hard-filter variants improve upon removing residual filtering, showing that directional partitioning is useful even with binary weights. However, R-CDFR improves AutoAttack over the hard-threshold control by 1.57 points on CIFAR-10 and 1.92 points on CIFAR-100. Relative to the rank-matched top-k control, the gains are 1.09 and 1.37 points, respectively.

R-CDFR also obtains lower S_P, S_R, G_P, and D_R, together with a larger ΔQ on both datasets. The larger suppression contrast indicates that continuous damping concentrates regularization more strongly toward low-gain directions while avoiding a discontinuous rank boundary. These results support the use of a smooth gain-conditioned split rather than a fixed binary partition.



(a) Clean and robust accuracy. (b) Filtered diagnostics. (c) Operator and solver behavior.

Figure 6: Sensitivity to the damping parameter λ on CIFAR-10. The horizontal axis is the multiplier relative to $\lambda_0 = 10^{-3}$. (a) Clean and AutoAttack accuracy. (b) S_P, S_R, and G_P, normalized by their values at λ_0. (c) Effective pass dimension and ten-iteration CG residual, also normalized by their values at λ_0. Error bars denote standard deviation over three seeds.

E.4 Damping Sensitivity

We study the effect of the damping parameter by setting $\lambda = m_\lambda \lambda_0$, where $\lambda_0 = 10^{-3}$ is the default value and

$$m_\lambda \in \{0.1, 0.3, 1, 3, 10\}.$$

All other objective terms, training settings, and evaluation protocols are held fixed. The sweep is conducted on CIFAR-10 with three independent seeds.

For each trained model, we additionally measure the normalized effective pass dimension

$$d_{\text{eff}}(\lambda) = \mathbb{E}_x \left[\frac{\operatorname{tr}(P_\lambda(x))}{\operatorname{rank}(J_h(x))} \right], \tag{24}$$

and the relative residual of the ten-iteration CG solve

$$r_{\text{CG}}(\lambda) = \mathbb{E} \left[\frac{\|A_\lambda w_{10} - g\|_2}{\|g\|_2} \right], \tag{25}$$

where w_{10} denotes the approximate solution after ten CG iterations. For the mechanism plots, each quantity is divided by its value at λ_0.

Figure 6 shows that AutoAttack accuracy is stable around the default damping value. Across $m_\lambda \in [0.3, 3]$, the maximum variation in AutoAttack accuracy is less than one percentage point, while clean accuracy remains nearly unchanged. Performance decreases at both ends of the tested range, indicating that neither an almost undamped residual nor an excessively residual-dominant allocation is preferred.

As λ increases, S_P and the effective pass dimension decrease, whereas S_R and G_P increase. This trend is consistent with the spectral responses in Eq. (5): larger damping places more weight on the residual filter and concentrates the pass-filtered perturbation on a smaller set of higher-gain directions. The CG residual decreases with λ, reflecting the improved conditioning of A_λ. Together, the results show that the default value lies in a regime that balances spectral allocation, finite-step consistency, and numerical accuracy.

F Additional Training Recipe and Architecture Evaluations

F.1 Matched TRADES-AWP recipe

All rows use WRN-34-10, 200 training epochs, cosine learning-rate decay, 10-step PGD training at $\epsilon = 8/255$, no extra data, and three independent seeds. EMA and CutMix are applied identically in the base and R-CDFR settings, so the comparison isolates the addition of the R-CDFR regularizers.

Dataset	Method	Clean	AA
CIFAR-10	TRADES-AWP + EMA/CutMix	82.67±0.28	58.02±0.34
CIFAR-10	**+ R-CDFR**	**82.79±0.25**	**59.31±0.37**
CIFAR-100	TRADES-AWP + EMA/CutMix	68.11±0.33	37.18±0.46
CIFAR-100	**+ R-CDFR**	**68.20±0.31**	**38.54±0.51**

Table 10: Incremental evaluation under a matched TRADES-AWP recipe with EMA and CutMix. Values are percentages over three seeds.

Adding R-CDFR improves AutoAttack by 1.29 points on CIFAR-10 and 1.36 points on CIFAR-100, while mean clean accuracy changes by $+0.12$ and $+0.09$ points.

F.2 DeiT-S

Method	C10 Clean	C10 AA	C100 Clean	C100 AA
AWP	79.92±0.31	51.08±0.43	64.72±0.42	29.84±0.58
TART	80.34±0.29	51.86±0.39	65.48±0.40	31.12±0.56
R-CDFR	**80.71±0.27**	**53.29±0.41**	**66.07±0.38**	**33.36±0.53**

Table 11: DeiT-S/4 results (percent; three seeds). The pre-logit class-token embedding is used as h_ψ.

R-CDFR improves DeiT-S AutoAttack accuracy by 1.43 points on CIFAR-10 and 2.24 on CIFAR-100 relative to TART, while also improving clean accuracy. The ordering across both datasets indicates that the regularizer transfers to a token-based representation without changing its operator definition.

F.3 ResNet-18

Method	CIFAR-10 Clean	CIFAR-10 AA	CIFAR-100 Clean	CIFAR-100 AA
Standard ERM	92.43±0.18	0.07±0.02	73.76±0.31	0.03±0.01
PGD-AT	60.82±0.46	45.03±0.72	**56.18±0.50**	26.54±0.88
AWP	69.46±0.43	43.77±0.69	50.84±0.48	27.41±0.84
TART	74.12±0.39	45.92±0.67	53.88±0.46	28.18±0.82
R-CDFR	**75.93±0.37**	**48.74±0.70**	54.83±0.44	**29.62±0.79**

Table 12: ResNet-18 results under the same ℓ_∞ evaluation protocol. Values are mean ± standard deviation over three independent runs. Bold denotes the best result among adversarially trained methods.

R-CDFR improves AutoAttack over TART by 2.82 points on CIFAR-10 and 1.44 on CIFAR-100 with ResNet-18. The same ordering on a lower-capacity backbone supports transfer beyond WRN-34-10.

G Directional Evaluation under Known Geometry

We evaluate four analytic circle and plane tasks for which the tangent space and its orthogonal complement are known independently of the learned representation (Zhang et al., 2022b). Tangent error is the mean principal angle to the analytic tangent space. For $D \in \{T, N\}$, $\widetilde{d}_D = \mathbb{E}[d_D/\epsilon_D]$, where d_D is the smallest classification-changing perturbation constrained to the corresponding analytic direction family and ϵ_D is its reference budget. Each entry reports the mean and standard deviation over three seeds after selecting the worst task separately for each seed. True-normal robust accuracy uses 20-step PGD with 10 random starts in the analytic normal space, with radius 0.10 on the circle tasks and 0.20 on the plane tasks.

	CIFAR-10 ($\epsilon_2 = 0.5$)			CIFAR-100 ($\epsilon_2 = 0.5$)			TinyImageNet ($\epsilon_2 = 1.0$)		
Method	Clean↑	PGD-ℓ_2↑	AA-ℓ_2↑	Clean↑	PGD-ℓ_2↑	AA-ℓ_2↑	Clean↑	PGD-ℓ_2↑	AA-ℓ_2↑
PGD-AT	71.92±0.78	69.80±0.47	62.86±0.83	58.79±0.52	51.28±0.90	48.91±0.95	53.97±0.53	39.16±0.97	36.02±1.05
TRADES	78.41±0.39	66.10±0.74	63.75±0.79	68.86±0.44	43.92±0.82	41.30±0.86	56.21±0.48	31.88±0.88	29.41±0.92
MART	74.06±0.43	72.84±0.73	70.92±0.78	64.31±0.50	52.47±0.89	49.86±0.93	55.08±0.51	40.21±0.94	36.88±0.99
AWP	73.55±0.75	71.83±0.51	66.99±0.82	61.17±0.46	54.26±0.85	51.72±0.88	54.62±0.47	41.02±0.91	37.91±0.95
TART	78.49±0.38	72.10±0.71	70.24±0.79	66.61±0.45	53.44±0.88	50.91±0.91	55.12±0.47	40.55±0.91	37.46±0.96
R-CDFR	**79.97±0.36**	**74.62±0.65**	**72.94±0.71**	67.09±0.44	**56.13±0.74**	**53.66±0.78**	54.48±0.43	**43.08±0.86**	**39.87±0.90**

Table 14: ℓ_2 **robustness evaluation.** R-CDFR exhibits consistent gains under a complementary threat model, indicating that its improvements are not confined to the ℓ_∞ evaluation. Values are mean $\pm$ standard deviation over five independent runs on CIFAR-10 and CIFAR-100 and three independent runs on TinyImageNet.

Method	Tangent error ↓	$\widetilde{d}_T$ ↑	$S_{N^\star}$ ↓	$\widetilde{d}_N$ ↑	True-normal RA ↑
TRADES	$(21.8\pm0.9)^\circ$	1.16±0.04	0.193±0.009	0.91±0.03	72.4±0.8
TART	$(14.6\pm0.7)^\circ$	1.34±0.04	0.144±0.007	1.11±0.04	78.1±0.7
R-CDFR (ambient GP)	$(13.9\pm0.7)^\circ$	1.41±0.04	0.118±0.006	1.18±0.04	80.5±0.7
R-CDFR w/o residual	$(12.8\pm0.6)^\circ$	1.48±0.04	0.126±0.006	1.19±0.04	80.2±0.7
R-CDFR	$\mathbf{(11.9\pm0.6)^\circ}$	**1.52±0.04**	**0.081±0.005**	**1.34±0.04**	**84.3±0.6**

Table 13: Results on analytic circle and plane tasks. $S_{N^\star} = \mathbb{E}[\|P_{N^\star} g\|_2 / \|g\|_2]$ uses the ground-truth normal projector. The ambient-GP row replaces $\mathcal{L}_{\text{res}}$ with an intensity-matched ambient gradient penalty.

R-CDFR obtains the smallest tangent error, the largest normalized decision distances in both analytic direction families, and the highest true-normal robust accuracy. The comparison with the residual ablation isolates the contribution of $\mathcal{L}_{\text{res}}$: adding the residual term raises true-normal robust accuracy from 80.2% to 84.3% and lowers $S_{N^\star}$ from 0.126 to 0.081. The ambient-gradient control reaches 80.5% true-normal robust accuracy, showing that the gain-selective residual penalty provides a stronger directional effect than intensity-matched uniform gradient shrinkage. Because the tangent and normal spaces are analytic in these tasks, this table provides the direct geometric counterpart to the model-relative natural-image diagnostics.

H Additional Robustness Evaluations

H.1 ℓ_2-bounded robustness

Motivation. The main paper focuses on ℓ_∞ robustness. Since R-CDFR conditions regularization on the learned representation spectrum rather than on a single ambient norm, it is important to verify that the improvements generalize across different threat models.

Protocol. We evaluate robustness using PGD-ℓ_2 and AutoAttack-ℓ_2. We set the perturbation budget to $\epsilon_2 = 0.5$ for CIFAR-10/100 and $\epsilon_2 = 1.0$ for TinyImageNet. PGD-ℓ_2 uses 20 steps with step size $\alpha = 2\epsilon_2/20$ and random initialization within the ℓ_2 ball. AutoAttack-ℓ_2 uses the default parameter-free configuration. All evaluations are performed on the test set. Clean accuracy is reported alongside robust accuracy for the evaluated checkpoints.

Results. Table 14 shows that R-CDFR consistently improves robust accuracy across datasets under both PGD-ℓ_2 and AutoAttack-ℓ_2. The robust-accuracy improvements are consistent across the three datasets. This indicates that the gains are not confined to the ℓ_∞ evaluation used in the main paper.

H.2 Fixed-reference pass/residual stress test

We call this diagnostic PR-20 because its directions are generated by finite pass and residual filters under a frozen reference encoder. The pass budget is $\tau_P = 0.50$ on CIFAR-10, 0.40 on CIFAR-100, and 0.30 on TinyImageNet, while the residual budget is $\varepsilon_R = 8/255$ on all three datasets. Each of the 20 iterations first updates the pass-filtered component and projects it onto $\{u_P : \|u_P\|_{A_\lambda} \le \tau_P\}$,

	CIFAR-10		CIFAR-100		TinyImageNet	
Method	Clean↑	PR-20↑	Clean↑	PR-20↑	Clean↑	PR-20↑
PGD-AT	62.86±0.47	45.12±0.79	58.79±0.52	26.48±0.92	53.97±0.53	22.31±0.98
TRADES	78.41±0.39	49.36±0.76	**68.86±0.44**	28.94±0.85	**56.21±0.48**	24.66±0.91
AWP	66.99±0.51	51.74±0.73	61.17±0.46	30.88±0.82	54.62±0.47	26.03±0.89
TART	78.49±0.38	53.21±0.71	66.61±0.45	32.04±0.80	55.12±0.47	27.86±0.87
R-CDFR	**79.97±0.36**	**55.84±0.69**	67.09±0.44	**34.91±0.77**	54.48±0.43	**30.12±0.84**

Table 15: Robust accuracy under PR-20, an external fixed-reference stress test that alternates pass-filtered and residual-filtered updates. Clean accuracy is included for context. Values are mean ± standard deviation over five independent runs on CIFAR-10 and CIFAR-100 and three independent runs on TinyImageNet. Bold denotes the best result in each column.

then updates the residual-filtered component and projects it onto $\{u_R : \|u_R\|_2 \leq \varepsilon_R\}$. The reference encoder, the damped metric, and both filters are held fixed throughout the attack.

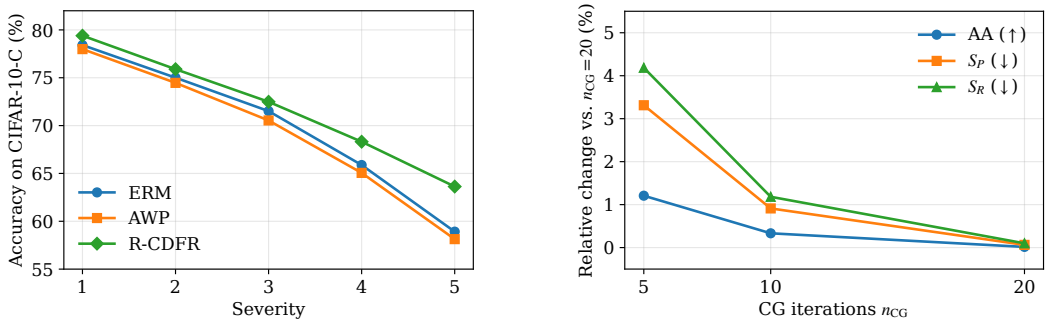
R-CDFR obtains the highest PR-20 accuracy on all three datasets. Relative to TART, the gains are 2.63, 2.87, and 2.26 points on CIFAR-10, CIFAR-100, and TinyImageNet. Because the encoder and both filters are frozen throughout the attack, this result tests the pass/residual perturbation family without allowing each evaluated model to redefine the attack geometry. The consistent ordering therefore complements the ambient AutoAttack result with a filter-aligned external stress test.

H.3 Common corruptions

	CIFAR-10-C		CIFAR-100-C		TinyImageNet-C	
Method	Acc.↑	mCE↓	Acc.↑	mCE↓	Acc.↑	mCE↓
Standard ERM	**74.10±0.29**	100.00±0.00	**51.20±0.38**	100.00±0.00	**40.30±0.42**	100.00±0.00
PGD-AT	70.85±0.35	92.30±0.76	49.02±0.43	95.60±0.88	39.10±0.47	96.40±0.94
TRADES	71.64±0.33	90.80±0.71	49.55±0.41	94.10±0.83	39.42±0.45	95.20±0.90
AWP	72.08±0.31	89.60±0.68	50.12±0.39	92.80±0.79	39.68±0.44	94.60±0.86
TART	72.31±0.30	88.90±0.66	50.26±0.38	92.10±0.77	39.74±0.43	94.20±0.84
R-CDFR	73.02±0.28	**86.70±0.62**	51.03±0.36	**89.30±0.72**	40.26±0.41	**92.80±0.81**

Table 16: Common-corruption accuracy and mean corruption error on CIFAR-10-C, CIFAR-100-C, and TinyImageNet-C. Values are mean ± standard deviation over five independent runs on CIFAR-10-C and CIFAR-100-C and three independent runs on TinyImageNet-C. The ERM mCE is the normalization reference and is therefore fixed at 100.00. Bold and underlined values denote the best and second-best results in each column, respectively.

R-CDFR obtains the lowest mCE on all three corruption benchmarks and the highest corruption accuracy among the adversarial-training and geometry-aware methods. ERM remains slightly higher in average corruption accuracy, while R-CDFR narrows the gap to 1.08, 0.17, and 0.04 points on CIFAR-10-C, CIFAR-100-C, and TinyImageNet-C. The severity curves in Figure 7a show that the advantage over AWP becomes larger at higher corruption levels, extending the observed stability beyond adversarially optimized perturbations.



(a) Corruption accuracy vs. severity.

(b) CG convergence sensitivity.

Figure 7: External and numerical diagnostics. (a) CIFAR-10-C accuracy across corruption severity. (b) Relative change in AutoAttack and filtered-sensitivity metrics as the number of CG iterations increases.

Figure 7b shows that the relative changes in AutoAttack and the two filtered-sensitivity scores shrink sharply between five and ten iterations and are negligible by twenty iterations. The numerical trend agrees with Table 20.

I Representation-Conditioned Controls

I.1 Fixed-reference controls

Method	CIFAR-10 Clean	CIFAR-10 AA	CIFAR-100 Clean	CIFAR-100 AA
TART (joint)	78.49±0.38	55.38±0.79	66.61±0.45	33.20±0.91
R-CDFR (joint)	**79.97±0.36**	**57.43±0.71**	**67.09±0.44**	**36.98±0.78**
TART (fixed h)	77.91±0.41	54.64±0.81	65.92±0.47	32.47±0.93
R-CDFR (fixed h)	**79.21±0.39**	**56.58±0.74**	**66.41±0.46**	**35.76±0.82**

Table 17: Fixed-representation control. The representation h is pretrained and held fixed while the classifier and regularization terms are optimized. Values are mean ± standard deviation over five independent runs. Bold denotes the better result within each representation setting.

Holding h fixed lowers absolute performance for both methods, but R-CDFR retains a 1.94-point AutoAttack advantage on CIFAR-10 and a 3.29-point advantage on CIFAR-100. The gain therefore persists when the representation geometry is stationary and only the classifier and regularization objective are optimized.

	CIFAR-10			CIFAR-100		
Method	$S_P^{\mathrm{ref}}\downarrow$	$S_R^{\mathrm{ref}}\downarrow$	$G_P^{\mathrm{ref}}\downarrow$	$S_P^{\mathrm{ref}}\downarrow$	$S_R^{\mathrm{ref}}\downarrow$	$G_P^{\mathrm{ref}}\downarrow$
AWP	1.35±0.07	1.94±0.05	1.10±0.04	2.08±0.11	3.05±0.08	1.80±0.06
TART	1.32±0.05	1.57±0.07	0.96±0.02	2.03±0.06	2.43±0.10	1.58±0.04
R-CDFR	**1.11±0.04**	**1.29±0.03**	**0.78±0.05**	**1.69±0.07**	**1.95±0.06**	**1.19±0.03**

Table 18: Cross-representation diagnostics computed using a separate frozen ERM reference encoder. All methods are evaluated under the same external representation geometry. Values are mean ± standard deviation over five independent runs. Lower is better.

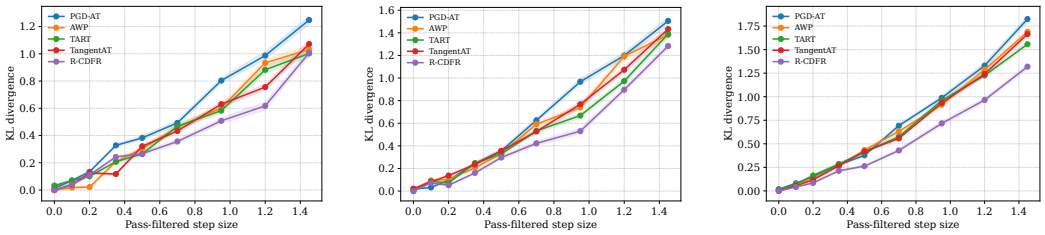
Under the separate ERM reference encoder, R-CDFR remains lowest on all six diagnostic entries. Relative to TART, it reduces S_P^{ref}, S_R^{ref}, and G_P^{ref} by 15.9%, 17.8%, and 18.8% on CIFAR-10 and by

16.7%, 19.8%, and 24.7% on CIFAR-100. The smaller margins than in Table 2 are consistent with evaluating every method under a common external geometry, while the preserved ordering shows that the diagnostic advantage is not confined to each model's own Jacobian spectrum.

J Finite-Step Consistency and Representation Visualization

J.1 Pass-filtered consistency curves

Figure 8 evaluates predictive consistency as the step size ε_P in Eq. (11) increases. Each perturbation is formed by filtering and normalizing a Gaussian input direction through P_λ, and the vertical axis reports the KL divergence between the clean and perturbed predictions.



(a) CIFAR-10 (b) CIFAR-100 (c) TinyImageNet

Figure 8: Predictive KL divergence as the pass-filtered step size ε_P increases, where $\delta_P = \varepsilon_P P_\lambda u/(\|P_\lambda u\|_2 + \eta)$. Lower curves indicate stronger finite-step predictive consistency along pass-filtered perturbations.

R-CDFR follows the lowest or near-lowest KL trajectory over the full step range on all three datasets. The separation becomes clearer at moderate and large steps, which is precisely the regime targeted by the finite-displacement KL term rather than the local scalar slope alone. The curve-level result agrees with the ablation in Table 3, where removing $\mathcal{L}_{\text{cons}}$ produces the largest increase in G_P while leaving the residual-energy measure close to the full model.

J.2 TinyImageNet Filtered Diagnostics

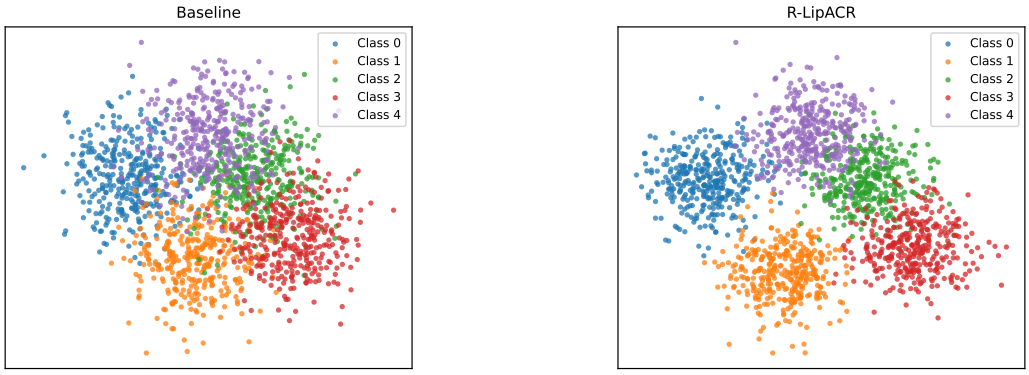
Table 19 reports the TinyImageNet counterpart of the model-relative diagnostics in Table 2. The evaluation protocol and damping parameter are unchanged.

J.3 Embedding-space visualization

Figure 9 visualizes five CIFAR-10 classes using the same sample set and UMAP configuration for TRADES and R-CDFR.

Method	$S_P \downarrow$	$S_R \downarrow$	$G_P \downarrow$
PGD-AT	2.85±0.10	4.22±0.14	2.36±0.09
TRADES	3.27±0.11	4.74±0.15	2.61±0.10
MART	2.71±0.09	3.98±0.13	2.18±0.08
AWP	2.46±0.09	3.71±0.12	2.03±0.08
Parseval	4.96±0.17	6.73±0.21	3.74±0.13
Jacobian Reg.	4.31±0.15	6.12±0.19	3.28±0.12
Lipschitz-Margin	3.93±0.14	5.78±0.18	3.02±0.11
TangentAT	2.53±0.09	2.95±0.10	1.78±0.07
TART	2.41±0.08	2.83±0.10	1.69±0.07
Consistency AT	2.44±0.09	2.88±0.10	1.63±0.07
MAT	2.62±0.09	3.08±0.11	1.86±0.08
AE-Proj	3.08±0.10	3.64±0.12	2.27±0.09
R-CDFR	**1.86±0.06**	**2.21±0.08**	**1.28±0.05**

Table 19: Model-relative diagnostics on TinyImageNet. Definitions and evaluation settings match Table 2. Values are mean ± standard deviation over three independent runs. Lower is better.



(a) TRADES baseline

(b) R-CDFR

Figure 9: UMAP visualization of penultimate-layer CIFAR-10 embeddings for five classes. The same UMAP configuration and sample set are used in both panels. The figure is used as a qualitative check of class structure under robust training.

The R-CDFR embedding exhibits more compact within-class groups and clearer separation for several class pairs, while substantial overlap remains. The visualization complements the quantitative robustness and consistency results by showing that the regularizer preserves a structured feature space.

K Numerical Stability

K.1 CG accuracy

	CIFAR-10			CIFAR-100		
n_{CG}	AA	S_P	S_R	AA	S_P	S_R
5	56.92	0.944	1.119	36.41	1.505	1.774
10	57.39	0.919	1.092	36.87	1.476	1.727
20	57.43	0.915	1.084	36.98	1.469	1.714

Table 20: Sensitivity to CG iterations. The default 10-iteration solve is close to the 20-iteration result.

Increasing CG iterations from five to ten changes AutoAttack by less than half a point and substantially reduces the discrepancy in both filtered-sensitivity scores. The twenty-iteration result is nearly saturated, supporting ten iterations as a practical accuracy–cost compromise. Figure 7b shows the same convergence trend in relative terms.