TOWARDS SHARED FLAT REGIONS: BOOSTING ADVERSARIAL TRANSFERABILITY BY EXPONENTIALLY TILTING CURVATURE-REGULARIZED GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer-based adversarial attacks rely on accessible surrogate models to fool unseen targets, but often overfit surrogate-specific loss landscapes. Although flatness-aware attacks mitigate this issue, local flatness on a surrogate does not guarantee agreement with target models. We investigate neighborhood gradient stability as a surrogate-observable signal for constructing transferable attack directions, motivated by the hypothesis that effective perturbations exploit stable, high-loss regions shared across models. We propose Tilting Curvature Regularized Attack (TCRA), which combines response-guided finite-difference probing, parallel–orthogonal gradient decomposition, and exponential tilting. Local probes capture directional gradient changes, while tilting prioritizes candidate directions from high-loss neighborhoods. Our analysis bounds the deviation of the constructed direction and establishes sufficient conditions for target-loss ascent under model discrepancy, including momentum and projected sign updates. Experiments on ImageNet-compatible and CIFAR-10 benchmarks demonstrate improved transferability across standard, adversarially trained, and Transformer-based targets. TCRA also improves the evaluated gradient-based and input-transformation attacks without target queries or surrogate-parameter tuning.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable success in various applications, yet they remain highly vulnerable to adversarial examples (Goodfellow et al., 2015; Dong et al., 2018). These imperceptible perturbations can induce incorrect predictions while remaining visually indistinguishable, raising critical security concerns in safety-sensitive applications such as autonomous driving (Evtimov et al., 2017) and facial authentication systems (Sharif et al., 2016). A fundamental property that enables practical adversarial attacks is *transferability* (Papernot et al., 2016; Dong et al., 2018), where adversarial examples generated from a surrogate model can successfully fool unknown target models. This property makes transfer-based attacks a powerful black-box attack paradigm, as attackers can exploit only accessible surrogate models without requiring knowledge of target architectures or parameters.

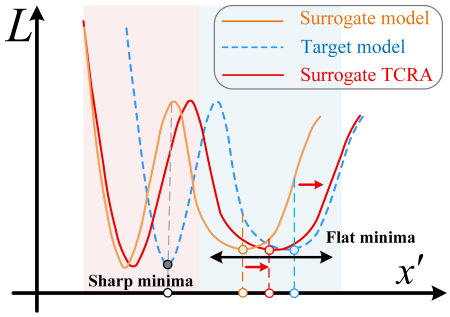


Figure 1: Insight of TCRA. Existing flatness attacks seek surrogate-specific flat regions, whereas TCRA identifies curvature-stable directions that approximate flat regions shared target models.

Despite extensive progress, achieving highly transferable adversarial examples remains challenging. Existing studies mainly attribute limited transferability to the overfitting problem: adversarial perturbations optimized on a surrogate model may exploit model-specific decision boundaries and fail to generalize to unseen models (Liu et al., 2017b; Wu et al., 2020b). To alleviate this issue, previous approaches have explored improved optimization strategies, including momentum-based gradient accumulation (Dong et al., 2018), Nesterov accelerated updates (Lin et al., 2019), variance reduc-

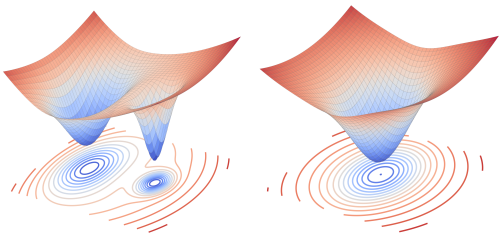
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

tion (Wu et al., 2021), and input transformation techniques (Xie et al., 2019; Wang et al., 2021a). Although these methods improve transferability, they mainly focus on obtaining more stable gradients or diverse input distributions, while the underlying geometric property governing cross-model transfer remains insufficiently explored.

Recent studies have revealed an important connection between flat loss landscapes and adversarial transferability (Zhang et al., 2022; Zhao et al., 2022). The intuition is that perturbations located in flat regions are less sensitive to parameter variations and therefore exhibit better generalization. However, existing flatness-aware attacks implicitly optimize flatness only with respect to the surrogate model. We argue that this assumption is fundamentally incomplete: *a perturbation that lies in a flat region of the surrogate model does not necessarily transfer well, since the corresponding region may become sharp or unstable in an unseen target model.* As illustrated in Figure 1, the flatness of a surrogate model does not necessarily imply cross-model generalization. A region that appears flat on the surrogate landscape may correspond to an unstable or sharp region on the target model due to landscape discrepancies. Therefore, directly optimizing surrogate-specific flatness can still lead to model-dependent perturbations. Instead, we hypothesize that highly transferable adversarial examples should lie in ***model-shared flat regions***, where the local loss landscape remains smooth across both surrogate and target models, as illustrated in Figure 2. Such regions capture model-agnostic directions that are insensitive to architectural and parameter differences, providing a more reliable mechanism for cross-model generalization. However, directly optimizing shared flatness is challenging because target models are unavailable in black-box attack scenarios. This raises a question:

How can we discover adversarial perturbations residing in target-shared flat regions using only surrogate-model information?

To address this challenge, we propose Tilting Curvature Regularized Attack (TCRA), a novel framework that implicitly explores model-shared flat regions through local curvature regularization. Instead of explicitly estimating the Hessian matrix, which is computationally prohibitive for large-scale models, TCRA employs neighborhood-based finite differences to approximate curvature-aware gradient information. Specifically, we decompose local gradient variations into dominant and orthogonal components, where the dominant directions capture task-specific sharp variations while orthogonal components provide more stable cross-model directions. By applying exponential tilting, TCRA adaptively emphasizes directions with high adversarial effectiveness while suppressing unstable curvature-sensitive components.



(a) Misaligned Flat (b) Shared Flat (ours)

Figure 2: (a) Misaligned flat regions: Different models converge to distinct flat basins, causing optimization conflicts. (b) Shared flat basin: Models converge to a common low-loss region, enabling better generalization.

Different from previous attacks that optimize flatness of a single surrogate model, TCRA aims to identify perturbation trajectories that are robust under local landscape variations, thereby approximating shared flat regions between surrogate and unknown target models. Moreover, TCRA introduces no additional model training or target-model queries and can be seamlessly integrated with existing momentum and transformation-based attack frameworks.

Our main contributions are summarized as follows:

• We reveal a new perspective on adversarial transferability: successful transferable adversarial examples should reside in *model-shared flat regions* rather than merely flat regions of the surrogate model. Based on this insight, we formulate transferability from the perspective of cross-model landscape generalization.

• We propose TCRA, a novel curvature-regularized transfer attack that implicitly explores shared flat regions without accessing target models. TCRA efficiently approximates local curvature through finite differences and leverages structured gradient decomposition with exponential tilting to obtain stable transferable perturbation directions.

• Extensive experiments on ImageNet-compatible benchmarks demonstrate that TCRA achieves superior transferability against both standard and adversarially trained models. Furthermore, TCRA is

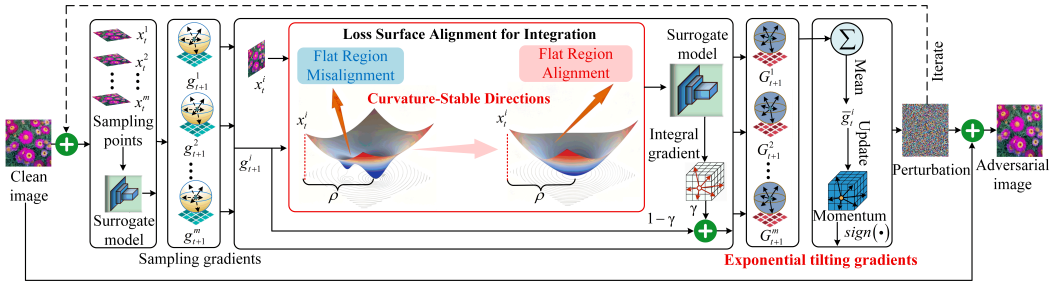


Figure 3: Overall framework of TCRA. TCRA explores model-shared flat regions by estimating curvature-stable directions from surrogate models, and generates transferable adversarial examples through exponential tilting of curvature-aware gradients.

compatible with existing gradient-based and input-transformation attacks, consistently improving their transfer performance.

2 Related Work

Transfer-based adversarial attacks. Transfer-based attacks craft adversarial examples on accessible surrogates without querying target models, distinguishing them from query-based black-box attacks (Brendel et al., 2017; Wang et al., 2022). Optimization-based methods stabilize update directions through momentum (MI-FGSM) (Dong et al., 2018), Nesterov acceleration (NI-FGSM) (Lin et al., 2019), and neighborhood-based gradient refinement (VMI-FGSM and EMI-FGSM) (Wang & He, 2021; Wang et al., 2021b). Complementary approaches promote transferability through input transformations and gradient aggregation, including DIM (Xie et al., 2019), TIM (Dong et al., 2019), SIM (Lin et al., 2019), Admix (Wang et al., 2021a), and SSA (Long et al., 2022). Ensemble attacks exploit shared vulnerabilities across accessible models (Liu et al., 2017a; Chen et al., 2024). Our method complements these strategies by explicitly probing local gradient changes and separating their directional contributions.

Landscape-aware transfer attacks. Recent methods connect transferability to adversarial loss geometry. PGN (Ge et al., 2023) penalizes gradient norms to seek flat adversarial maxima, while CWA (Chen et al., 2024) studies common weaknesses through flatness and proximity to optima across ensemble models. AWT (Chen et al., 2025) improves adversarial landscapes through surrogate-parameter tuning; LPAA (Mi et al., 2025) explores local perturbation subspaces; and PFA (Tao et al., 2026) extends flatness considerations to the optimization path. These studies motivate a distinction between *local stability on accessible models* and *agreement with unseen targets*: the former does not by itself guarantee the latter. TCRA investigates local gradient variation as a surrogate-observable signal for transferable direction construction. Its distinctive mechanism combines response-guided probes, orthogonal gradient decomposition, and loss-based exponential aggregation, without target queries or surrogate-parameter tuning. Model-shared flatness serves as a motivating hypothesis, rather than a property certified from surrogate information.

Flatness and gradient regularization. Flat minima have long been studied in relation to generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Neyshabur et al., 2017; Jiang et al., 2019). SAM (Foret et al., 2021) optimizes neighborhood loss, whereas gradient-norm regularization (Zhao et al., 2022) provides an alternative route toward flatter solutions. These ideas motivate adversarial landscape regularization, but parameter-space generalization does not directly establish input-space transferability. We focus on neighborhood gradient variation: a small gradient at one point need not imply that nearby gradients remain consistent.

Exponential tilting. Exponential tilting controls emphasis placed on individual losses through exponential reweighting (Li et al., 2023; Aminian et al., 2024). In TCRA, it aggregates geometrically constructed directions across sampled neighborhoods, interpolating between uniform averaging and emphasis on high-loss samples. The two mechanisms play distinct roles: local probes determine candidate directions, while tilting determines their relative contributions. Contribution lies in coupling directional geometry with loss-dependent aggregation, rather than in exponential weighting.

3 METHODOLOGY

Preliminaries. Let $x \in [0,1]^d$ be a clean image with label y, and let $J_s(z) = J(z, y; \theta_s)$ denote the loss of an accessible surrogate model. We consider untargeted transfer attacks under an ℓ_∞ perturbation budget ϵ. Writing $\mathcal{X}_\epsilon(x) = \{z \in [0,1]^d : \|z - x\|_\infty \le \epsilon\}$, the surrogate attack objective is $\max_{z \in \mathcal{X}_\epsilon(x)} J_s(z)$. The resulting example is evaluated on unseen target models without using their gradients, parameters, or predictions during generation. Throughout, $g(z) = \nabla J_s(z)$ denotes the input gradient. Maximizing the surrogate loss promotes misclassification on the surrogate; transfer additionally requires that the resulting perturbation exploit vulnerabilities shared with other models.

3.1 BEYOND POINTWISE FLATNESS: STABILITY OF THE LOCAL GRADIENT FIELD

Flatness does not guarantee transferability. Flatness-aware attacks seek regions insensitive to input perturbations (Qin et al., 2022; Ge et al., 2023), but surrogate stability does not imply cross-model agreement. Even a zero-curvature surrogate loss can have gradients opposing those of a target. Transfer therefore requires locally stable directions that remain effective across models.

A surrogate-observable stability measure. We characterize local gradient variation by

$$\mathcal{C}_\zeta(z) = \max_{\|u\|_2 \le \zeta} \|g(z+u) - g(z)\|_2. \tag{1}$$

Here the surrogate loss is assumed to be defined on the probe neighborhood. This quantity measures changes in the gradient field, including both magnitude and direction. In particular, a small gradient at one point need not imply a stable neighborhood. If the Hessian $H_s = \nabla^2 J_s$ is ρ-Lipschitz on this neighborhood, then $\left|\mathcal{C}_\zeta(z) - \zeta \|H_s(z)\|_2\right| \le \frac{\rho}{2}\zeta^2$. Thus, at sufficiently small radii, gradient variation measures local curvature without conflating it with gradient magnitude. It does not measure variation of the Hessian itself.

What stability can imply across models. A conditional bound makes the missing cross-model requirement explicit. Let J_q be a target loss satisfying $\|\nabla J_q(z) - g(z)\|_2 \le b$ at the current point. For any $\|u\|_2 \le \zeta$ with $g(z+u) \neq 0$, define $v_u = g(z+u)/\|g(z+u)\|_2$. Then

$$\langle \nabla J_q(z), v_u \rangle \ge \|g(z+u)\|_2 - b - \mathcal{C}_\zeta(z). \tag{2}$$

The bound follows by adding and subtracting $g(z+u)$ and applying Cauchy–Schwarz. It shows that local gradient variation consumes part of the margin available for cross-model ascent. Lower variation strengthens this bound when gradient strength and model discrepancy are controlled. It does not remove the discrepancy term b, which is unavailable during a transfer attack, or establish that every low-curvature example transfers better. Moreover, the bound concerns a normalized local gradient, not the complete projected, momentum-based update below.

Design principle. We balance adversarial effectiveness and neighborhood stability through

$$\max_{z \in \mathcal{X}_\epsilon(x)} \{J_s(z) - \lambda \mathcal{C}_\zeta(z)\}, \qquad \lambda \ge 0. \tag{3}$$

We hypothesize that controlling gradient variation promotes transfer when surrogate and target gradients share useful structure. Here, *model-shared flat regions* denote stable, high-adversarial-loss neighborhoods across models, whose shared geometry cannot be verified from the surrogate alone. To avoid nested optimization, `TCRA` combines finite-difference probing, parallel–orthogonal decomposition, and loss-based exponential tilting. The resulting update is motivated by Eq. equation 3, rather than its exact gradient.

3.2 TILTING CURVATURE-REGULARIZED ATTACK

Paired finite-difference probes. At iteration t, sample N independent offsets $u_i \sim \mathrm{Unif}([-\zeta, \zeta]^d)$ and set $z_i = x_t^{adv} + u_i$. This specifies the computational sampling neighborhood separately from the Euclidean neighborhood used in Eq. 1. A box offset can have Euclidean norm as large as $\sqrt{d}\,\zeta$; the small-radius interpretation in Eq. 3.1 should therefore not be applied to these samples with an unchanged radius. Probe points are auxiliary loss evaluations, whereas the final attack iterate is projected onto $\mathcal{X}_\epsilon(x)$. For clarity, omit the sample index temporarily and define $n_1(a) = a/\|a\|_1$ for $a \neq 0$, with $n_1(0) = 0$. Using a probe step $s > 0$, evaluate

$$g_0 = g(z), \qquad z_1 = z - s\, n_1(g_0), \qquad g_1 = g(z_1). \tag{4}$$

The difference $r = g_1 - g_0$ records the gradient response to a local descent probe. We use this response to select a second probe center and repeat the evaluation:

$$z_2 = z - s\, n_1(r), \qquad g_2 = g(z_2), \qquad r = g_1 - g_0, \tag{5}$$

$$z_3 = z_2 - s\, n_1(g_2), \qquad g_3 = g(z_3). \tag{6}$$

These negative-direction probes inspect the neighborhood; the final adversarial update remains an ascent step. Four input-gradient evaluations are required per sample. The finite-difference interpretation is standard. If H_s is ρ-Lipschitz on the segment between z and $z + sv$, then

$$\left\| \frac{g(z + sv) - g(z)}{s} - H_s(z)v \right\|_2 \leq \frac{\rho s}{2} \|v\|_2^2. \tag{7}$$

In particular, r/s approximates $-H_s(z)n_1(g_0)$. The paired differences reveal curvature along selected directions without forming a Hessian. They neither estimate its full spectrum nor solve the maximization defining $\mathcal{C}_\zeta$. The error bound applies on smooth segments; finite differences remain computable when this smoothness condition does not hold.

Separate gradient strength from directional change. For $\delta \in [0, 1]$, combine the gradients at the two centers and their corresponding probe points:

$$h_+ = \delta g_0 + (1 - \delta) g_2, \qquad h_- = \delta g_1 + (1 - \delta) g_3. \tag{8}$$

The subscripts distinguish center and probe mixtures; they do not label proven stable and unstable components. Let $P_+ = h_+ h_+^\top / \|h_+\|_2^2$ when $h_+ \neq 0$, and let $P_+ = 0$ otherwise. Decompose

$$h_p = P_+ h_-, \qquad h_o = (I - P_+) h_-, \qquad d = \delta h_+ + (1 - \delta) h_o. \tag{9}$$

Because $(I - P_+)h_+ = 0$, the orthogonal residual satisfies $h_o = (I - P_+)(h_- - h_+)$. It therefore isolates the part of the probe-induced change that is not parallel to the center mixture. The update retains the center direction and introduces a separately weighted transverse correction, rather than accumulating the parallel response twice. Orthogonality alone does not establish that this correction is more transferable; its utility is an empirical component of the method. For $h_+ \neq 0$ and $0 < \delta \leq 1$, the construction has the exact properties

$$\langle d, h_+ \rangle = \delta \|h_+\|_2^2, \qquad \|d\|_2^2 = \delta^2 \|h_+\|_2^2 + (1 - \delta)^2 \|h_o\|_2^2. \tag{10}$$

Thus, the transverse correction preserves a positive component along h_+ while adjusting the direction. This is a geometric property relative to the center mixture, not a guarantee of target-loss ascent. We retain the shared coefficient δ in Eqs. 8 and 9 to match Algorithm 1.

Loss-based exponential tilting. Restore the sample index and let $\ell_i = J_s(z_i)$. Aggregate the sample directions using $w_i = \frac{\exp(c(\ell_i - \ell_{\max}))}{\sum_{j=1}^N \exp(c(\ell_j - \ell_{\max}))}$, $\bar{d}_t = \sum_{i=1}^N w_i d_i$, where $c \geq 0$ and $\ell_{\max} = \max_j \ell_j$. Subtracting $\ell_{\max}$ improves numerical stability without changing the weights. Uniform averaging is recovered at $c = 0$; larger c concentrates weight on higher-loss neighborhoods. For $c > 0$, these weights solve the entropy-regularized selection problem

$$w = \arg\max_{a \in \Delta_N} \left\{ \sum_{i=1}^N a_i \ell_i - \frac{1}{c} \mathrm{KL}(a \| \pi) \right\}, \tag{11}$$

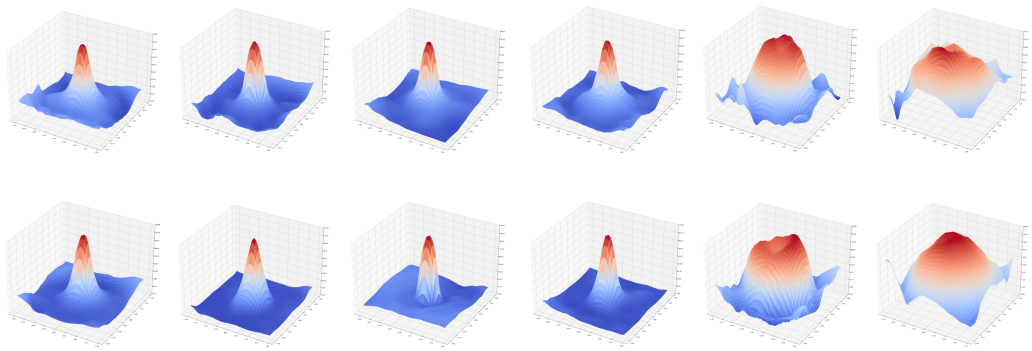
where Δ_N is the probability simplex and $\pi_i = 1/N$. Tilting therefore trades loss preference against concentration relative to uniform sampling. It does not directly select low-curvature samples or guarantee variance reduction. Curvature information enters through d_i; adversarial effectiveness enters through w_i. Because d_i is a constructed direction rather than $\nabla \ell_i$, Eq. equation 3.2 is not generally the gradient of a log-sum-exp loss.

Momentum and feasibility. Starting from $m_0 = 0$ and $x_0^{adv} = x$, use the aggregated direction in the standard momentum update:

$$m_{t+1} = \mu m_t + n_1(\bar{d}_t), \qquad x_{t+1}^{adv} = \Pi_{\mathcal{X}_\epsilon(x)} \left(x_t^{adv} + \eta\, \mathrm{sign}(m_{t+1}) \right). \tag{12}$$

Return x_T^{adv}, with $s = \eta = \epsilon / T$ and optional momentum ($\mu = 0$ disables it). Each iteration requires $4N$ surrogate gradient evaluations, without explicit Hessian storage or target queries. Computational comparisons should report gradient counts and runtime.

Methodological distinction. `TCRA` combines response-guided probing, parallel–orthogonal decomposition, and loss-based neighborhood aggregation. Local geometry shapes candidate directions, while exponential tilting weights their contributions. Cross-model sharing remains an empirical hypothesis evaluated on held-out targets.



(a) MI-FGSM (b) NI-FGSM (c) VMI-FGSM (d) EMI-FGSM (e) PGN (f) TCRA (ours)

Figure 4: Visualization of loss surfaces along two random directions for two randomly sampled adversarial examples on the surrogate model. The center of each 2D graph corresponds to the adversarial example generated by different attack methods.

4 Supporting Analysis for TCRA

Finite-difference error. Assume J_s is twice continuously differentiable and its Hessian is ρ-Lipschitz on the relevant segment. The fundamental theorem of calculus gives

$$\frac{g(z+sv)-g(z)}{s} - H_s(z)v = \int_0^1 [H_s(z+\tau sv) - H_s(z)]v \, d\tau. \tag{13}$$

Taking norms and integrating $\rho\tau s\|v\|_2^2$ proves Eq. 7. This result justifies the directional probe interpretation, not an equality between the complete TCRA direction and a regularized objective gradient.

Local curvature interpretation. Taylor's formula yields $g(z+u) - g(z) = H_s(z)u + R(u)$ with $\|R(u)\|_2 \le \rho\|u\|_2^2/2$. Consequently, $\mathcal{C}_\zeta(z) \le \zeta\|H_s(z)\|_2 + \rho\zeta^2/2$. For a unit right singular vector v attaining $\|H_s(z)v\|_2 = \|H_s(z)\|_2$, evaluating at $u = \zeta v$ gives the matching lower bound. Together these establish Eq. 3.1.

Conditional cross-model ascent. For the direction v_u defined in the main text, Cauchy–Schwarz gives $\langle \nabla J_q(z), v_u\rangle \ge \|g(z+u)\|_2 - \|\nabla J_q(z) - g(z+u)\|_2$. The triangle inequality bounds the last term by $b + \mathcal{C}_\zeta(z)$, proving Eq. equation 2. If J_q is also L_q-smooth along the step and $a = \|g(z+u)\|_2 - b - \mathcal{C}_\zeta(z) > 0$, then $J_q(z+\tau v_u) - J_q(z) \ge \tau a - \frac{L_q}{2}\tau^2 > 0, 0 < \tau < 2a/L_q$. Here $L_q > 0$, and the statement concerns the unprojected step along v_u. It does not automatically extend to the sign, projection, decomposition, or momentum operations of TCRA, nor does a loss increase alone imply misclassification.

Why a discrepancy condition is necessary. Consider $J_s(z) = z$ and $J_q(z) = -z$ on a bounded one-dimensional interval. Both losses have zero Hessian and zero gradient variation, but every positive surrogate-ascent step decreases the target loss. Constants may be added to make both losses nonnegative without changing this conclusion. This example rules out a universal transfer guarantee based only on surrogate curvature.

Exponential tilting. For $c > 0$, differentiating the simplex-constrained objective in Eq. equation 11 at its interior optimum yields $a_i \propto \pi_i \exp(c\ell_i)$. Strict concavity gives the unique optimum, recovering Eq. equation 3.2. As $c \to 0$, the weights approach $1/N$; as $c \to \infty$, they concentrate uniformly on the maximal-loss samples. Neither limit imposes a curvature preference.

5 Theoretical Analysis

We analyze how TCRA's local direction construction, momentum, and exponential tilting affect cross-model ascent. The analysis provides sufficient conditions for increasing an unseen target loss under explicit surrogate–target discrepancy assumptions. Proofs are provided in Appendices C–E.

Table 1: The untargeted attack success rates (%±std, over 10 random runs) of various gradient-based attacks in the single model setting. The adversarial examples are crafted on Inc-v3, Inc-v4, IncRes-v2, and Res-101 by MI-FGSM (MI), NI-FGSM (NI), VMI-FGSM (VMI), EMI-FGSM (EMI), RAP, and our TCRA, respectively. Here * indicates the white-box model.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	MI	**100.0±0.0***	51.0±0.47	45.8±0.60	49.0±0.24	22.6±0.52	22.0±0.35	10.9±0.24
	NI	**100.0±0.0***	61.4±0.42	59.6±0.54	57.2±0.18	22.5±0.37	22.7±0.35	11.5±0.26
	VMI	**100.0±0.0***	74.8±0.58	69.9±0.92	65.5±0.67	41.6±0.54	41.6±0.54	25.0±0.34
	EMI	**100.0±0.0***	80.7±0.58	77.1±0.37	72.4±0.83	33.0±0.61	31.9±0.49	17.0±0.48
	RAP	99.9±0.10*	84.5±0.69	79.3±0.47	76.5±0.65	56.9±0.84	51.3±0.62	31.9±0.35
	PGN	100.0±0.0*	90.6±0.67	89.5±0.75	81.2±0.68	64.6±0.75	65.6±0.94	45.3±0.77
	TCRA	**100.0±0.0***	**96.1±0.57**	**94.3±0.72**	**86.5±0.55**	**70.1±0.72**	**70.9±0.92**	**51.2±0.55**
Inc-v4	MI	57.2±0.36	**100.0±0.0***	46.1±0.14	51.5±0.33	19.1±0.46	18.4±0.23	10.2±0.36
	NI	62.8±0.43	**100.0±0.0***	52.7±0.34	56.7±0.19	19.2±0.25	18.3±0.37	11.7±0.29
	VMI	77.6±0.65	99.8±0.10*	69.8±0.41	66.7±0.33	41.1±0.87	41.2±0.54	27.0±0.24
	EMI	84.2±0.62	99.7±0.10*	75.0±0.70	74.4±0.64	31.5±0.44	28.0±0.65	16.2±0.36
	RAP	85.3±0.74	99.5±0.21*	79.5±0.62	77.2±0.42	45.2±0.69	46.8±0.48	29.3±0.51
	PGN	91.2±0.58	99.6±0.15*	87.6±0.74	83.5±0.53	67.0±0.68	64.2±0.63	49.1±0.82
	TCRA	**97.1±0.52**	99.8±0.32*	**93.2±0.57**	**89.0±0.33**	**72.9±0.38**	**70.3±0.54**	**55.1±0.28**
IncRes-v2	MI	58.2±0.21	52.4±0.41	99.3±0.21*	50.7±0.26	22.0±0.37	22.0±0.31	13.8±0.43
	NI	60.3±0.35	57.1±0.17	99.5±0.17*	55.3±0.35	18.3±0.18	19.3±0.29	12.1±0.16
	VMI	78.2±0.64	77.0±0.57	99.1±0.36*	66.0±0.48	47.6±0.69	43.3±0.36	37.7±0.37
	EMI	85.2±0.78	83.3±0.29	99.7±0.18*	74.0±0.56	38.4±0.48	33.8±0.53	24.1±0.48
	RAP	87.1±0.75	84.2±0.45	99.4±0.28*	79.4±0.64	50.3±0.47	49.8±0.89	40.2±0.54
	PGN	92.0±0.69	92.3±0.63	**99.8±0.10***	83.5±0.41	74.6±0.75	71.5±0.64	66.62±0.58
	TCRA	**96.1±0.52**	**97.0±0.37**	99.7±0.12*	**89.0±0.29**	**79.9±0.82**	**76.6±0.55**	**70.9±0.62**
Res-101	MI	51.5±0.26	42.2±0.35	36.3±0.24	**100.0±0.0***	18.7±0.32	16.6±0.14	9.0±0.22
	NI	55.6±0.35	46.9±0.41	40.8±0.28	**100.0±0.0***	17.5±0.57	17.6±0.42	9.2±0.24
	VMI	75.0±0.40	69.2±0.59	63.0±0.84	**100.0±0.0***	35.9±0.41	35.7±0.87	24.1±0.57
	EMI	74.3±0.65	71.7±0.47	62.6±0.29	**100.0±0.0***	25.7±0.74	24.6±0.98	13.3±0.68
	RAP	80.4±0.75	75.5±0.56	68.0±0.84	99.9±0.10*	40.3±0.47	39.9±0.73	30.4±1.03
	PGN	86.2±0.84	83.3±0.66	77.8±0.69	100.0±0.0*	63.1±1.32	62.9±0.74	50.8±0.88
	TCRA	**91.9±0.32**	**88.6±0.33**	**83.7±0.51**	**100.0±0.0***	**69.0±0.83**	**69.1±0.64**	**55.5±0.77**

Table 2: The untargeted attack success rates (%) of various gradient-based attacks in the multi-model setting. The adversarial examples are generated on the ensemble models, *i.e.* Inc-v3, Inc-v4, and IncRes-v2. Here * indicates the white-box model.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Avg.
MI	99.8*	99.5*	97.8*	66.8	68.4	36.7	36.9	22.6	66.06
NI	**100.0***	**99.9***	99.6*	74.6	74.0	37.7	37.1	22.2	68.14
VMI	99.9*	99.5*	98.0*	82.2	81.6	66.8	64.1	50.8	80.36
EMI	**100.0***	**99.9***	**99.7***	92.3	93.0	62.8	59.5	40.0	80.90
RAP	**100.0***	99.4*	98.2*	93.5	93.4	70.1	69.8	58.9	85.41
PGN	**100.0***	**99.9***	99.6*	94.2	94.6	88.2	86.6	79.2	92.78
TCRA	**100.0***	**99.9***	99.5*	**95.5**	**95.7**	**93.1**	**92.0**	**84.3**	**95.00**

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTINGS

Dataset. We use a standard ImageNet-compatible dataset containing 1,000 images of size $299 \times 299 \times 3$ with ground-truth and target labels for targeted attack evaluation. **Models.** Experiments are conducted on five standard pre-trained models: Inc-v3, Inc-v4, IncRes-v2, Res-101, and Res-152, along with three adversarially trained models: Inc-v3ens3, Inc-v4ens4, and IncRes-v2ens.

Baselines. We compare `TCRA` with six advanced attacks: MI-FGSM, NI-FGSM, VMI-FGSM, EMI-FGSM, RAP, and PGN. To evaluate `TCRA`'s generality, we also combine it with input transformations including DIM, TIM, SIM, Admix, SSA, and STM. Since our experimental setup follows the same configuration as the PGN (Ge et al., 2023).

Hyper-parameters. We set $\epsilon = 16/255$, $T = 10$, and $\alpha = \epsilon/T$. For MI-FGSM and NI-FGSM, the decay factor is $\mu = 1.0$. For VMI-FGSM, we use $N = 20$ samples with neighborhood size

Table 3: The untargeted attack success rates (%) of our TCRA method, when it is integrated with DIM, TIM, SIM, Admix, and SSA, respectively. The adversarial examples are generated on Inc-v3. Here * indicates the white-box model.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Avg.
DIM	99.7*	72.2	67.3	62.8	32.8	30.7	16.4	54.56
TCRA-DIM	**100.0***	**96.3**	**95.2**	**92.4**	**78.3**	**83.2**	**59.8**	**86.46**
TIM	99.9*	51.6	47.2	47.8	29.6	30.7	20.5	46.76
TCRA-TIM	**100.0***	**92.5**	**89.5**	**81.2**	**82.9**	**83.4**	**71.7**	**85.89**
SIM	**100.0***	70.5	68.2	63.8	37.5	37.8	22.0	57.11
TCRA-SIM	**100.0***	**95.5**	**95.2**	**88.7**	**81.8**	**80.3**	**65.1**	**86.66**
Admix	**100.0***	78.6	77.3	69.5	41.6	40.3	24.1	61.63
TCRA-Admix	**100.0***	**95.2**	**95.6**	**90.2**	**82.5**	**83.1**	**65.5**	**87.44**
SSA	99.7*	88.3	86.8	77.7	56.7	55.3	35.2	71.39
TCRA-SSA	**99.9***	**93.3**	**94.3**	**87.7**	**75.2**	**74.3**	**55.5**	**82.89**

$\beta = 1.5 \times \epsilon$. EMI-FGSM adopts $N = 11$, interval bound $\eta = 7$, and linear sampling. RAP is configured with $\alpha = 2/255$, $K = 400$, inner loop $T = 10$, late-start $K_{LS} = 100$, and neighborhood size $\epsilon_n = 16/255$. For PGN and TCRA, we set the number of examples $N = 20$, the balanced coefficient $\delta = 0.5$, and the upper bound of $\zeta = 3.0 \times \epsilon$ to ensure fair runtime comparison.

6.2 Results

Loss Surfaces for Adversarial Example. In order to verify that the proposed TCRA method can help the adversarial samples to find flat localized regions, surface shapes of the loss functions where adversarial samples are generated using different methods are compared in this section, and the agent model used is Inc-v3 in Figure 4.

Attack a Single Model. We evaluate TCRA under the single-model setting against six representative gradient-based baselines, including MI-FGSM, NI-FGSM, VMI-FGSM, EMI-FGSM, RAP, and PGN. The attack success rate (ASR) is adopted as the evaluation metric, measuring the probability that generated adversarial examples successfully fool the target models. Following previous transfer-based attack evaluations, we generate adversarial examples from four surrogate models, including Inc-v3, Inc-v4, IncRes-v2, and Res-101. As shown in Table 1, TCRA consistently achieves superior transferability across both standard and adversarially trained target models. For instance, when adversarial examples are generated on Inc-v3 and transferred to Inc-v4, TCRA achieves an ASR of 96.1%, outperforming the strongest baseline RAP by 11.6 percentage points. Moreover, TCRA maintains substantial improvements on adversarially trained models, demonstrating its robustness against diverse target architectures. These results validate that explicitly optimizing curvature-stable directions enables adversarial perturbations to better generalize across different model landscapes.

Attack an Ensemble of Models. To further evaluate the cross-model transferability of TCRA, we conduct ensemble attacks following (Dong et al., 2018), where logits from Inc-v3, Inc-v4, and IncRes-v2 are averaged with equal weights. The generated adversarial examples are subsequently evaluated on both normally trained and adversarially trained target models. As shown in Table 2, TCRA achieves the highest black-box ASR of 95.00%, outperforming VMI-FGSM, EMI-FGSM, RAP, and PGN by 14.64, 14.10, 9.59, and 2.22 percentage points, respectively. Moreover, TCRA obtains consistent gains on adversarially trained models, surpassing PGN and RAP by 5.1 and 23.5 percentage points. These results demonstrate that modeling curvature-stable directions provides stronger cross-model generalization than optimizing surrogate flatness alone.

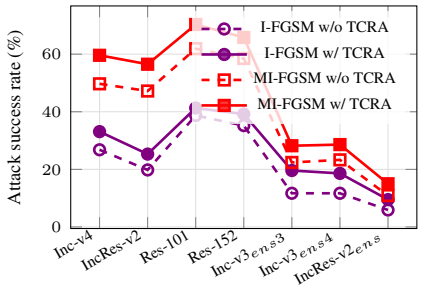


Figure 5: Average attack success rates of I-FGSM and MI-FGSM with and without TCRA on seven black-box models. Adversarial examples are generated on Inc-v3.

Combined with Input Transformation Attacks. To further demonstrate the generality of TCRA, we integrate it with five representative input transformation attacks, including DIM, TIM, SIM, Admix, and SSA. Adversarial examples are generated on Inc-v3 and evaluated on six black-box models. As shown in Table 3, TCRA consistently improves existing transformation-based attacks. For example, TCRA+DIM achieves an ASR of 86.46%, improving DIM by 31.90 percentage points. These consistent gains demonstrate that curvature-stable direction discovery provides complementary benefits beyond input diversity and enables TCRA to enhance diverse attack frameworks.

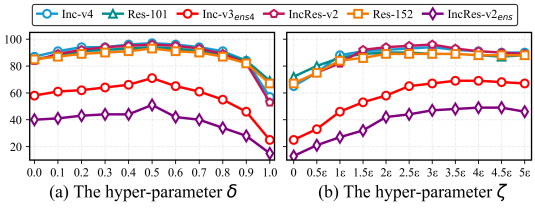
6.3 Ablation Study

Component Analysis. Table 4 verifies the effectiveness of each component in TCRA. Removing curvature estimation results in the largest performance degradation, decreasing the average ASR from 86.5% to 83.1% and increasing curvature variation from 0.126 to 0.184. This confirms that explicitly modeling local landscape variations is crucial for discovering curvature-stable adversarial directions. Gradient decomposition and exponential tilting further reduce curvature variation and improve transferability by suppressing unstable components and emphasizing robust adversarial directions. The full TCRA achieves the lowest curvature variation of 0.126 and the highest average ASR of 86.5%, demonstrating that all components contribute complementarily to locating flatter and more transferable adversarial regions.

Table 4: Ablation study of TCRA. ASR and curvature variation are averaged over six black-box target models. Lower curvature variation indicates greater curvature stability.

Method	ASR (%) ↑	Curvature Variation ↓
w/o Curvature Estimation	83.1	0.184
w/o Gradient Decomposition	83.4	0.171
w/o Exponential Tilting	84.2	0.158
TCRA	**86.5**	**0.126**

Effect of the Balanced Coefficient δ. The coefficient δ controls the balance between dominant and orthogonal curvature-aware directions. As shown in Figure. 6, transferability consistently improves with increasing δ and reaches the best performance at $\delta = 0.5$. This indicates that effectively combining stable principal directions and complementary orthogonal directions is critical for discovering transferable perturbation trajectories. We therefore set $\delta = 0.5$.



(a) The hyper-parameter δ (b) The hyper-parameter ζ

Figure 6: Untargeted attack success rates (%) on six black-box models under varying δ or ζ. Adversarial examples are generated by TCRA on Inc-v3.

Effect of the ζ-ball Radius. The radius ζ determines the neighborhood range for finite-difference curvature estimation. A small ζ provides insufficient landscape information, while an overly large neighborhood introduces irrelevant variations. As shown in Figure 6, $\zeta = 3.0 \times \epsilon$ achieves the best trade-off between curvature stability and transferability across different target models, adopted in all experiments.

7 Conclusion

In this work, we revisit adversarial transferability from the perspective of cross-model landscape generalization. We reveal that transferable adversarial examples should not merely reside in surrogate-specific flat regions, but follow curvature-stable directions that remain consistent across different models. Based on this insight, we propose TCRA, a curvature-aware framework that implicitly discovers model-shared flat regions through surrogate-observable gradient variations. By combining curvature decomposition with exponential tilting, TCRA promotes stable transferable directions while suppressing model-specific variations. Extensive experiments demonstrate consistent improvements in black-box transferability across diverse standard and adversarially trained models, highlighting curvature stability as a fundamental principle for transferable adversarial optimization.

REFERENCES

Gholamali Aminian, Amir R Asadi, Tian Li, Ahmad Beirami, Gesine Reinert, and Samuel N Cohen. Generalization error of the tilted empirical risk. *arXiv preprint arXiv:2409.19431*, 2024.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, 2024.

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*, 2025.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *In Proceedings of International Conference on Learning Representations*, 2021.

Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *In Proceedings of International Conference on Learning Representations*, 2021.

Zhijin Ge, Hongying Liu, Wang Xiaosen, Fanhua Shang, and Yuanyuan Liu. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36:70141–70161, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of International Conference on Learning Representations*, 2015.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *Proceedings of International Conference on Learning Representations*, 2018.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11936–11945, 2021.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6084–6092, 2019.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *Proceedings of International Conference on Learning Representations*, 2019.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of International Conference on Learning Representations*, 2017a.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *Proceedings of the European Conference on Computer Vision*, pp. 549–566. Springer, 2022.

Jian-Xun Mi, Xuanhui Zhong, and Weisheng Li. Improving adversarial transferability with local perturbation augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 262–271, 2020.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *In Proceedings of Advances in Neural Information Processing Systems*, 30, 2017.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. In *arXiv*, 2016.

Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *arXiv preprint arXiv:2210.05968*, 2022.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM on Asia Conference on Computer and Communications Security*, pp. 1528–1540, 2016.

Zeze Tao, Jinjia Peng, and Huibing Wang. Prompting adversarial transferability via path flatness attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16158–16167, 2021a.

Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *In Proceedings of the British Machine Vision Conference*, 2021b.

Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *Proceedings of The European Conference on Computer Vision (ECCV)*, pp. 156–174. Springer, 2022.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *Proceedings of International Conference on Learning Representations*, 2020a.

Weibin Wu, Yuxin Su, Michael R. Chen, Shuhui Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Towards understanding and improving the transferability of adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1706–1715, 2020b.

Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the transferability of adversarial samples with adversarial transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9024–9033, 2021.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *Proceedings of International Conference on Learning Representations*, 2018.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *In Proceedings of Network and Distributed System Security Symposium*, 2018.

Jian Zhang, Chen Huang, et al. Boosting adversarial transferability by achieving flat local maxima. In *NeurIPS*, 2022.

Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. In *Proceedings of International Conference on Machine Learning*, pp. 26982–26992. PMLR, 2022.

ETHICS STATEMENT

This work studies transferable adversarial attacks using existing ImageNet-compatible and CIFAR-10 benchmarks. TCRA aims to support robustness evaluation by identifying vulnerabilities that transfer across models without querying target models during adversarial example generation. However, improved attack transferability also poses dual-use risks, as the method could be misused to compromise deployed machine learning systems. The reported experiments concern benchmark evaluations and do not establish the feasibility or consequences of attacks in real-world deployments. We advocate using this method only for authorized security evaluation and research on improving model robustness.

REPRODUCIBILITY STATEMENT

The paper describes the TCRA algorithm, including its finite-difference probes, gradient decomposition, and exponential weighting. The theoretical assumptions, derivations, and proofs are provided in the main text and appendices. The experimental section specifies the benchmark dataset, surrogate and target models, baseline attacks, perturbation budgets, and attack hyperparameters. The appendices provide additional evaluations on Transformer-based targets, defense models, and CIFAR-10, along with computational cost analyses. Table 1 reports means and standard deviations over ten random runs for the single-model comparison. We will publicly release the implementation and experiment scripts upon acceptance.

AI USE STATEMENT

Large language models were used to assist with language editing, manuscript organization, and drafting supplementary material. They also assisted in identifying potential inconsistencies, examining mathematical formulations, and suggesting additional experiments and evaluation protocols. AI-generated content does not constitute independent evidence for theoretical or empirical claims. The authors are responsible for verifying all incorporated material and for the accuracy and integrity of the final manuscript, including its proofs, experimental results, and citations.

A ALGORITHM AND MATHEMATICAL SETUP

We study untargeted transfer attacks on a fixed surrogate loss $J_s(z) = J(z, y; \theta_s)$. Let $g(z) = \nabla J_s(z)$, $H(z) = \nabla^2 J_s(z)$, and $\mathcal{X} = \{z \in [0,1]^d : \|z - x\|_\infty \le \epsilon\}$. The label y and model parameters remain fixed. All gradients are with respect to the input. We distinguish the attack step η from the probe step s, even when both equal ϵ/T.

For a vector a, define $n_1(a) = a/\|a\|_1$ if $a \neq 0$ and $n_1(0) = 0$. The normalization is per image over all input coordinates. For $a \neq 0$, write $P_a = aa^\top/\|a\|_2^2$; set $P_0 = 0$. These definitions specify the zero-vector cases without introducing a division by zero.

At iterate x_t^{adv}, sample $u_i \sim \mathrm{Unif}([-\zeta, \zeta]^d)$ independently and set $z_i = x_t^{adv} + u_i$. Auxiliary probes are evaluated on the surrogate's input extension; only the attack iterate is projected onto $\mathcal{X}$. The statements below concern these unclipped probes. If auxiliary probes are clipped, their realized displacements must replace the prescribed displacements in the analysis.

Algorithm 1 stores a loss and a direction for each sample and performs normalized exponential aggregation once, after sampling. It requires four surrogate input-gradient evaluations per sample. The loss ℓ_i can be obtained from the forward evaluation used to compute $g_{i,0}$. No target-model evaluation is used in generation.

Assumptions and scope. The geometric bounds require an L-Lipschitz surrogate gradient on the segments connecting the relevant points. The finite-difference remainder bounds additionally require $J_s \in C^2$ and a ρ-Lipschitz Hessian there. Smoothness assumptions are local; they are not asserted globally for networks with nonsmooth activations. Target-loss statements explicitly assume target smoothness and a surrogate–target gradient discrepancy bound. These conditions are analytical conditions, not quantities supplied to the attack.

Algorithm 1 Tilting Curvature-Regularized Attack (TCRA)

Input: Clean input x, label y, surrogate loss J_s, budget ϵ
Input: Iterations T, samples N, steps $s, \eta > 0$, radius $\zeta \geq 0$
Input: Mixing coefficient $\delta \in [0,1]$, tilt $c \geq 0$, momentum $\mu \in [0,1]$

$x_0^{adv} \leftarrow x$, $m_0 \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  for $i = 1, \ldots, N$ **do**
    Sample $u_i \sim \mathrm{Unif}([-\zeta, \zeta]^d)$; set $z_i \leftarrow x_t^{adv} + u_i$
    $g_{i,0} \leftarrow g(z_i)$; $\ell_i \leftarrow J_s(z_i)$
    $z_{i,1} \leftarrow z_i - s n_1(g_{i,0})$; $g_{i,1} \leftarrow g(z_{i,1})$
    $r_i \leftarrow g_{i,1} - g_{i,0}$
    $z_{i,2} \leftarrow z_i - s n_1(r_i)$; $g_{i,2} \leftarrow g(z_{i,2})$
    $z_{i,3} \leftarrow z_{i,2} - s n_1(g_{i,2})$; $g_{i,3} \leftarrow g(z_{i,3})$
    $h_{i,+} \leftarrow \delta g_{i,0} + (1-\delta) g_{i,2}$
    $h_{i,-} \leftarrow \delta g_{i,1} + (1-\delta) g_{i,3}$
    $h_{i,o} \leftarrow h_{i,-} - P_{h_{i,+}} h_{i,-}$
    $d_i \leftarrow \delta h_{i,+} + (1-\delta) h_{i,o}$
  end for
  $\ell_{\max} \leftarrow \max_i \ell_i$; $w_i \leftarrow e^{c(\ell_i - \ell_{\max})} / \sum_j e^{c(\ell_j - \ell_{\max})}$
  $\bar{d}_t \leftarrow \sum_{i=1}^N w_i d_i$
  $m_{t+1} \leftarrow \mu m_t + n_1(\bar{d}_t)$
  $x_{t+1}^{adv} \leftarrow \Pi_{\mathcal{X}}(x_t^{adv} + \eta \operatorname{sign}(m_{t+1}))$
end for
return x_T^{adv}

B Local Gradient Variation and Finite Differences

For a Euclidean radius $R \geq 0$, define

$$\mathcal{C}_R(z) = \max_{\|u\|_2 \leq R} \|g(z+u) - g(z)\|_2. \tag{14}$$

A box sample with coordinate radius ζ satisfies $\|u_i\|_2 \leq R = \sqrt{d}\,\zeta$. Thus $\mathcal{C}_R$, not $\mathcal{C}_\zeta$, bounds all such samples. A large coordinate sampling radius does not automatically fall within the small-radius Taylor regime.

Lemma B.1 (Finite-difference error)**.** *If the Hessian is ρ-Lipschitz on $[z, z+sv]$, then, for $s > 0$,*

$$\left\| \frac{g(z+sv) - g(z)}{s} - H(z)v \right\|_2 \leq \frac{\rho s}{2} \|v\|_2^2. \tag{15}$$

Proof. The fundamental theorem of calculus gives

$$\frac{g(z+sv) - g(z)}{s} - H(z)v = \int_0^1 [H(z+\tau sv) - H(z)]v \, d\tau. \tag{16}$$

The integrand has norm at most $\rho\tau s\|v\|_2^2$. Integration proves the claim. In particular, taking $v = -n_1(g(z))$ interprets the first probe difference as a directional Hessian estimate. Since $\|n_1(g(z))\|_2 \leq 1$, its error is at most $\rho s/2$. □

Proposition B.2 (Curvature interpretation)**.** *If the Hessian is ρ-Lipschitz on the Euclidean ball of radius R about z, then*

$$\left| \mathcal{C}_R(z) - R\|H(z)\|_2 \right| \leq \tfrac{1}{2}\rho R^2. \tag{17}$$

Proof. For every $\|u\|_2 \leq R$, integration along $[z, z+u]$ yields $g(z+u) - g(z) = H(z)u + e(u)$ with $\|e(u)\|_2 \leq \rho\|u\|_2^2/2$. Taking the supremum gives the upper bound. For the lower bound choose a unit right singular vector v satisfying $\|H(z)v\|_2 = \|H(z)\|_2$ and evaluate at $u = Rv$. The reverse triangle inequality gives $\mathcal{C}_R(z) \geq R\|H(z)\|_2 - \rho R^2/2$. □

Gradient magnitude is a different quantity. For $J_s(z) = a^\top z$, $\mathcal{C}_R(z) = 0$ although $\|g(z)\|_2 = \|a\|_2$ can be arbitrarily large. Conversely, for $J_s(z) = K\|z\|_2^2/2$, the gradient at the origin is zero while $\mathcal{C}_R(0) = KR$. Neither a small pointwise gradient norm nor local gradient variation determines the other without additional assumptions.

Proposition B.3 (Derivative of the conceptual penalty)**.** *Suppose $J_s \in C^2$ on an open set containing the relevant balls. At z, suppose Eq. equation 14 has a unique maximizer u_* and $r_* = g(z + u_*) - g(z) \neq 0$. Then*

$$\nabla \mathcal{C}_R(z) = [H(z + u_*) - H(z)]^\top \frac{r_*}{\|r_*\|_2}. \tag{18}$$

Consequently, the gradient of $J_s(z) - \lambda \mathcal{C}_R(z)$ is $g(z) - \lambda \nabla \mathcal{C}_R(z)$ at this point.

Proof. The feasible offset ball is compact and independent of z. The objective is continuous. Uniqueness and the nonzero maximizing residual allow the envelope theorem to differentiate the maximizing value by differentiating the objective at fixed u_*. The Jacobian of $g(z + u_*) - g(z)$ is $H(z+u_*) - H(z)$. Applying the chain rule to its Euclidean norm proves the formula. No derivative of the optimizer u_* is needed. At nonunique maximizers or zero residuals the penalty need not be differentiable, and the displayed formula is not asserted. □

Equation equation 18 distinguishes the conceptual penalty from a gradient-norm penalty. Algorithm 1 does not solve the inner maximization or compute this derivative. Its probes and decomposition define a curvature-informed direction construction rather than an exact gradient method for the conceptual objective.

Finite precision. If the two computed gradients have Euclidean errors bounded by ε_g, the error relative to $H(z)v$ is bounded by

$$\frac{\rho s}{2}\|v\|_2^2 + \frac{2\varepsilon_g}{s}. \tag{19}$$

This follows from Lemma B.1 and the triangle inequality. Decreasing s controls truncation but can amplify numerical error. The bound concerns a fixed probe vector; errors in a data-dependent normalized vector require separate control. In particular, normalizing a nearly zero probe difference can be sensitive to roundoff.

C Geometry of the Constructed Direction

For one sample, abbreviate $a = h_+$, $b = h_-$, $o = (I - P_a)b$, and $d = \delta a + (1 - \delta)o$.

Proposition C.1 (Orthogonal residual and angular control)**.** *The residual satisfies $o = (I - P_a)(b - a)$ and $\langle a, o\rangle = 0$. Hence*

$$\langle d, a\rangle = \delta\|a\|_2^2, \qquad \|d\|_2^2 = \delta^2\|a\|_2^2 + (1 - \delta)^2\|o\|_2^2. \tag{20}$$

If $a \neq 0$ and $\delta > 0$, then

$$\cos\angle(d, a) = \left[1 + \left(\frac{(1 - \delta)\|o\|_2}{\delta\|a\|_2}\right)^2\right]^{-1/2}. \tag{21}$$

If the surrogate gradient is L-Lipschitz on both probe segments, $\|o\|_2 \leq Ls$.

Proof. The projector is orthogonal and annihilates the residual of a: $(I - P_a)a = 0$, including at $a = 0$. This proves the first identity and orthogonality. Expansion of the inner product and squared norm gives Eq. equation 20; division by $\|d\|_2\|a\|_2$ gives Eq. equation 21. Finally,

$$b - a = \delta(g_1 - g_0) + (1 - \delta)(g_3 - g_2). \tag{22}$$

Both probe displacements have norm at most s, so $\|b - a\|_2 \leq Ls$. An orthogonal projection has operator norm at most one, giving $\|o\|_2 \leq Ls$. □

This result controls the size of the transverse correction and its angle relative to the center mixture. It does not assign an intrinsically transferable meaning to orthogonality. The sign and usefulness of the transverse correction must be assessed empirically.

Theorem C.2 (Deviation of the aggregated direction)**.** *Fix an iterate x_t^{adv} and write $G_t = g(x_t^{adv})$. Suppose g is L-Lipschitz on the segments joining x_t^{adv} to z_i, z_i to $z_{i,2}$, and the two probe pairs. Let $R_i = \|z_i - x_t^{adv}\|_2$. For any nonnegative weights summing to one,*

$$\|\bar{d}_t - \delta G_t\|_2 \le \delta L \sum_i w_i R_i + (1-\delta^2) Ls. \tag{23}$$

For $\delta > 0$, set $q_t = \bar{d}_t/\delta$. Under box sampling,

$$\|q_t - G_t\|_2 \le E_t, \qquad E_t = L\sqrt{d}\,\zeta + \frac{1-\delta^2}{\delta} Ls. \tag{24}$$

Proof. The two center gradients obey $\|g_{i,0} - G_t\|_2 \le LR_i$ and $\|g_{i,2} - G_t\|_2 \le L(R_i + s)$. Therefore

$$\|h_{i,+} - G_t\|_2 \le LR_i + (1-\delta)Ls. \tag{25}$$

Combining this with Proposition C.1 yields

$$\|d_i - \delta G_t\|_2 \le \delta LR_i + (1-\delta^2)Ls. \tag{26}$$

Convex aggregation and the triangle inequality prove Eq. equation 23. Dividing by δ and using $R_i \le \sqrt{d}\,\zeta$ proves Eq. equation 24. □

The bound holds for the realized exponential weights, without independence between a weight and its direction. It is conservative and does not imply that tilting improves the bound, that smaller neighborhoods maximize empirical transfer, or that TCRA dominates an unmodified gradient attack. In particular, the comparison is unavailable at $\delta = 0$, when the retained center component disappears.

D Conditional Cross-Model Guarantees

Proposition D.1 (Local-gradient alignment)**.** *Suppose $\|\nabla J_q(z) - g(z)\|_2 \le B$. For $\|u\|_2 \le R$ with $g(z+u) \ne 0$, let $v = g(z+u)/\|g(z+u)\|_2$. Then*

$$\langle \nabla J_q(z), v \rangle \ge \|g(z+u)\|_2 - B - \mathcal{C}_R(z). \tag{27}$$

Proof. Add and subtract $g(z+u)$ in the inner product. Cauchy–Schwarz gives a lower bound of $\|g(z+u)\|_2 - \|\nabla J_q(z) - g(z+u)\|_2$. The triangle inequality bounds the subtracted term by $B + \mathcal{C}_R(z)$. □

Corollary D.2 (Normalized constructed direction)**.** *Under Theorem C.2, suppose $\delta > 0$, $q_t \ne 0$, and $\|\nabla J_q(x_t^{adv}) - G_t\|_2 \le B_t$. Then*

$$\left\langle \nabla J_q(x_t^{adv}), \frac{q_t}{\|q_t\|_2} \right\rangle \ge \|q_t\|_2 - B_t - E_t. \tag{28}$$

Proof. The distance between q_t and the target gradient is at most $B_t + E_t$. Adding and subtracting q_t in the inner product proves the bound. □

The next result covers the sign and projection actually used by the algorithm, and permits momentum through its update vector.

Theorem D.3 (Projected sign step)**.** *Let $z \in \mathcal{X}$, $m \in \mathbb{R}^d$, and $p = \Pi_{\mathcal{X}}(z + \eta\,\mathrm{sign}(m)) - z$. Suppose J_q is L_q-smooth on $[z, z+p]$. For any $\gamma > 0$, let $e = \gamma m - \nabla J_q(z)$ and $\theta_k = |p_k|$. Then*

$$J_q(z+p) - J_q(z) \ge \sum_k \theta_k |\nabla_k J_q(z)| - 2\sum_k \theta_k |e_k| - \tfrac{L_q}{2}\|p\|_2^2. \tag{29}$$

Proof. The feasible set is a coordinate box. Projection can shorten a signed displacement but cannot reverse its sign, so $p_k = \theta_k \,\mathrm{sign}(m_k)$ and $0 \le \theta_k \le \eta$. For scalars a, b, with $\mathrm{sign}(0) = 0$, we have $a\,\mathrm{sign}(b) \ge |a| - 2|a-b|$. If signs agree this is immediate; if signs disagree, $|a-b| \ge |a|$, and if $b = 0$ the same bound remains valid. Apply this inequality with $a = \nabla_k J_q(z)$ and $b = \gamma m_k$, multiply by θ_k, and sum. The smoothness lower bound $J_q(z+p) \ge J_q(z) + \langle \nabla J_q(z), p \rangle - L_q\|p\|_2^2/2$ completes the proof. □

A positive right-hand side guarantees a one-step target-loss increase. Projection may remove all available motion at a boundary, and target loss increase is not sufficient to certify misclassification. Neither outcome is assumed away in Eq. equation 29.

Lemma D.4 (Normalization perturbation)**.** *If $b \neq 0$, then for any vector a,*

$$\|n_1(a) - n_1(b)\|_1 \leq \frac{2\|a-b\|_1}{\|b\|_1}. \tag{30}$$

Proof. For $a \neq 0$, add and subtract $a/\|b\|_1$. The resulting bound is $(|\|b\|_1 - \|a\|_1| + \|a-b\|_1)/\|b\|_1$, and the reverse triangle inequality applies. For $a = 0$, the left side equals one and the right side equals two. □

Theorem D.5 (Momentum discrepancy and target-loss change)**.** *Suppose $\delta > 0$ and the conclusions of Theorem C.2 hold at iterations $j = 0, \ldots, t$, with errors E_j. Assume $\|G_j\|_1 \geq G_* > 0$ and $\|G_j - G_t\|_2 \leq L\|x_j^{adv} - x_t^{adv}\|_2$. Let $\|\nabla J_q(x_t^{adv}) - G_t\|_2 \leq B_t$ and $Q_t = \|\nabla J_q(x_t^{adv})\|_1 > 0$. Define*

$$S_t = \sum_{j=0}^{t} \mu^{t-j}, \qquad \omega_{t,j} = \mu^{t-j}/S_t, \tag{31}$$

with the convention that $\mu = 0$ puts all weight on $j = t$. Set

$$D_t = \frac{2\sqrt{d}}{G_*}\left[B_t + \sum_{j=0}^{t} \omega_{t,j}\big(E_j + L\|x_j^{adv} - x_t^{adv}\|_2\big)\right]. \tag{32}$$

Then the algorithm's momentum satisfies

$$\left\| \frac{m_{t+1}}{S_t} - n_1(\nabla J_q(x_t^{adv})) \right\|_1 \leq D_t. \tag{33}$$

For the actual projected step $p_t = x_{t+1}^{adv} - x_t^{adv}$, if J_q is L_q-smooth along that step, then

$$J_q(x_{t+1}^{adv}) - J_q(x_t^{adv}) \geq V_t - 2\eta Q_t D_t - \tfrac{L_q}{2}\|p_t\|_2^2, \tag{34}$$

where $V_t = \sum_k |p_{t,k}|\,|\nabla_k J_q(x_t^{adv})|$.

Proof. Because $\delta > 0$, $n_1(\bar{d}_j) = n_1(q_j)$, including when both vanish. Unrolling the momentum recursion gives $m_{t+1}/S_t = \sum_j \omega_{t,j} n_1(q_j)$. By Lemma D.4 and $\|v\|_1 \leq \sqrt{d}\|v\|_2$,

$$\|n_1(q_j) - n_1(G_j)\|_1 \leq 2\sqrt{d}\, E_j/G_*. \tag{35}$$

Applying the same lemma to (G_j, G_t) bounds their normalized difference by $2\sqrt{d}L\|x_j^{adv} - x_t^{adv}\|_2/G_*$. Applying it with $a = \nabla J_q(x_t^{adv})$ and $b = G_t$ bounds the target–surrogate normalized difference by $2\sqrt{d}B_t/G_*$. Convex averaging and the triangle inequality prove Eq. equation 33.

In Theorem D.3, choose $\gamma = Q_t/S_t$. Equation equation 33 gives $\|\gamma m_{t+1} - \nabla J_q(x_t^{adv})\|_1 \leq Q_t D_t$. Since $|p_{t,k}| \leq \eta$, the weighted error in Eq. equation 29 is at most $\eta Q_t D_t$, establishing Eq. equation 34. □

This theorem isolates three obstacles to target ascent: current model discrepancy, local direction-construction error, and staleness of momentum. It covers the implemented sign and box projection, but its sufficient condition $V_t > 2\eta Q_t D_t + L_q\|p_t\|_2^2/2$ may be conservative or fail in practice. It is neither an unconditional convergence theorem nor a comparison proving superiority over other attacks.

Why surrogate flatness alone is insufficient. On a bounded scalar interval, take $J_s(z) = z$ and $J_q(z) = -z$. Both have zero gradient variation and Hessian, yet every positive surrogate-ascent step decreases the target loss. Adding constants makes both losses nonnegative without changing their gradients. Thus even perfect surrogate flatness does not remove the need for cross-model information in a transfer guarantee.

E Exponential Tilting: Exact Properties and Limitations

Fix the sampled losses $\ell_1, \ldots, \ell_N$ and directions $d_1, \ldots, d_N$. Let $w_i(c) = e^{c\ell_i} / \sum_j e^{c\ell_j}$, $\pi_i = 1/N$, and Δ_N be the probability simplex.

Proposition E.1 (Entropy-regularized selection)**.** *For $c > 0$, the unique maximizer of*

$$\max_{a \in \Delta_N} \left\{ \sum_i a_i \ell_i - \frac{1}{c} \mathrm{KL}(a \| \pi) \right\} \tag{36}$$

is $w(c)$, and its optimal value is $c^{-1} \log(N^{-1} \sum_i e^{c\ell_i})$.

Proof. Write $Z = \sum_i e^{c\ell_i}$. Direct substitution gives

$$\sum_i a_i \ell_i - \frac{1}{c} \mathrm{KL}(a \| \pi) = \frac{1}{c} \log(Z/N) - \frac{1}{c} \mathrm{KL}(a \| w). \tag{37}$$

The last term is nonpositive and vanishes precisely at $a = w$, proving both claims, including boundary points of the simplex under the usual convention $0 \log 0 = 0$. □

Proposition E.2 (Loss preference and directional covariance)**.** *For fixed samples, define $\bar{\ell} = \sum_i w_i \ell_i$ and $\bar{d} = \sum_i w_i d_i$. Then*

$$\frac{d\bar{\ell}}{dc} = \sum_i w_i (\ell_i - \bar{\ell})^2 \geq 0, \qquad \frac{d\bar{d}}{dc} = \sum_i w_i (\ell_i - \bar{\ell}) d_i. \tag{38}$$

For any fixed vector a, the derivative of $\langle a, \bar{d} \rangle$ equals the weighted covariance between ℓ_i and $\langle a, d_i \rangle$.

Proof. Differentiating the normalized exponential gives $w_i' = w_i(\ell_i - \bar{\ell})$. Substitution into the two weighted sums proves Eq. equation 38. Taking an inner product and using $\sum_i w_i(\ell_i - \bar{\ell}) = 0$ gives the covariance statement. □

Consequently, increasing c favors larger surrogate losses, but improves target alignment only when the relevant covariance is positive. At $c = 0$ the weights are uniform; as $c \to \infty$ they concentrate uniformly on the maximal-loss samples. With $N = 1$, $w_1 = 1$ and tilting has no effect. The diagnostic $N_{\text{eff}} = (\sum_i w_i^2)^{-1}$ lies in $[1, N]$; concentration need not reduce estimator variance. Neither these identities nor the variational representation implies preferential selection of low-curvature samples.

F Combination with Other Attack Families

F.1 Gradient-Based Attacks

We combine the direction construction with I-FGSM, MI-FGSM, NI-FGSM, VMI-FGSM, EMI-FGSM, and RAP. In the no-additional-neighborhood-sampling variant, the construction replaces the local direction at the baseline's evaluation point. When only one candidate is aggregated, the exponential weight equals one; this variant evaluates the probe and decomposition components rather than the contribution of tilting.

Table 5 reports the resulting ASRs. Its existing Avg. column includes the source model and is an all-model average. For RAP, the average over the six black-box columns rises from 63.62% to 73.02%, a gain of 9.40 percentage points. The corresponding all-model averages are 68.80% and 76.87%. These results support compatibility in the evaluated configurations without establishing compatibility with every gradient-based attack.

F.2 Skip Gradient Method

SGM (Wu et al., 2020a) modifies the backward computation, and can be combined with TCRA by applying the same modified backward rule at each probe. Table 6 reports results with ResNet-18 as the source. The mean ASR increases from 54.77% for SGM and 69.67% for TCRA to 71.17%

Table 5: Untargeted attack success rates (%) of our TCRA method, when it is integrated with I-FGSM (BIM), MI-FGSM, NI-FGSM, VMI-FGSM, EMI-FGSM, and RAP, respectively. The adversarial examples are generated on Inc-v3. * indicates the white-box model.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Avg.
BIM	**100.0***	28.1	20.6	26.7	11.9	11.9	5.0	29.17
TCRA-BIM	**100.0***	**36.9**	**28.6**	**33.7**	**16.5**	**17.3**	**9.9**	**34.70**
MI	**100.0***	50.8	46.3	48.9	23.3	22.2	11.7	43.31
TCRA-MI	**100.0***	**59.5**	**56.9**	**58.8**	**28.7**	**28.6**	**14.9**	**49.63**
NI	**100.0***	62.3	59.8	57.8	22.4	22.3	11.8	48.06
TCRA-NI	**100.0***	**71.9**	**68.8**	**65.2**	**29.2**	**29.5**	**15.9**	**54.36**
VMI	**100.0***	75.2	70.2	66.0	41.7	40.9	24.8	59.83
TCRA-VMI	**100.0***	**82.1**	**79.0**	**72.7**	**48.3**	**50.1**	**31.5**	**66.24**
EMI	**100.0***	81.6	77.0	72.0	32.8	32.2	17.5	59.01
TCRA-EMI	**100.0***	**86.6**	**84.6**	**79.3**	**38.3**	**37.4**	**21.5**	**63.96**
RAP	99.9	84.6	79.3	76.6	57.2	51.3	32.7	68.80
TCRA-RAP	**100.0***	**95.1**	**89.9**	**88.8**	**62.6**	**59.5**	**42.2**	**76.87**

for their combination. The improvement is an average effect: the Inc-v3$_{ens3}$ entry decreases from 36.7% for TCRA to 36.4% for TCRA-SGM.

The Hessian interpretation in Appendix B applies to true gradients of a smooth scalar loss. A modified backward field need not be the gradient of such a loss; in that case finite differences measure changes in that field, and the Hessian statements do not automatically apply. The projection and tilting identities remain algebraic properties of the constructed vectors.

Table 6: Untargeted attack success rates (%) of our TCRA method with SGM method in the single model setting. The adversarial examples are crafted on Res-18.

Source:Res-18	Res-101	Res-152	Inc-v3	Inc-v4	IncRes-v2	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	Avg.
MI	82.8	73.3	54.5	48.7	33.9	17.4	18.1	46.96
SGM	89.1	82.5	66.0	58.8	45.4	20.8	20.8	54.77
TCRA	**96.3**	**94.1**	**80.7**	**78.5**	**61.9**	**36.7**	**39.5**	**69.67**
TCRA-SGM	**97.5**	**95.1**	**83.6**	**79.5**	**66.1**	**36.4**	**40.0**	**71.17**

F.3 Query-Based Attacks

Query-based attacks optimize against a target using its output scores or decisions. TCRA instead uses accessible surrogate gradients and performs no target queries. Adapting its probes to estimated target gradients would require a separate query-budget analysis. Transfer examples can in principle initialize a query-based procedure, but such a hybrid is outside the experiments reported here.

G Computational Cost and Finite-Difference Evaluation

Gradient evaluations. Let C_g denote the cost of one surrogate input-gradient evaluation. TCRA uses $4NT$ such evaluations, with leading computational cost $O(NTC_g)$ and vector-operation cost $O(NTd)$. For $N = 20$ and $T = 10$, the count is 800 evaluations per image before additional transformations or model ensembling. Equal iteration and sample counts do not imply equal costs across methods. If a comparison method uses two gradients per sample, its corresponding count is $2NT$.

Hessian versus Hessian-vector products. An explicitly stored $d \times d$ Hessian requires $O(d^2)$ entries. An automatic-differentiation Hessian-vector product need not form that matrix and can use a constant number of differentiation passes, with implementation-dependent memory costs. Finite differences use two gradient evaluations per product, or one additional evaluation if the base gradient is cached. Thus finite differences avoid explicit Hessian construction, but do not universally dominate automatic-differentiation HVPs in runtime or accuracy.

Table 7: Comparison of the approximation effect between directly optimizing the Hessian matrix (H_m) and using the Finite Difference Method (FDM) to approximate. "Time" represents the total running time on 1,000 images, and "Memory" represents the computing memory size.

Attack	H_m	FDM	Inc-v3	Inc-v4	IncRes-v2	Res-101	Res-152	Time (s)	Memory (MiB)
I-FGSM	✗	✗	**100.0***	27.8	19.1	38.1	35.2	52.31	1631
	✓	✗	**100.0***	**39.2**	**30.2**	**47.0**	**45.5**	**469.54**	**7887**
	✓	✓	**100.0***	37.9	28.6	45.7	44.6	96.42	1631

Stored quantities. Storing every d_i requires $O(Nd)$ vector memory. A streaming implementation can instead maintain a running maximum loss, exponential normalizer, and rescaled weighted direction sum, reducing aggregation memory to $O(d)$. Neither count includes surrogate activations or parameters.

Finite-difference comparison. Table 7 reports the supplied I-FGSM-based comparison on $1{,}000$ images using an RTX 2080 Ti, excluding data loading. The entries labeled H_m and FDM have times of 469.54 and 96.42 seconds and memory use of $7{,}887$ and $1{,}631$ MiB, respectively. The FDM configuration has slightly lower ASR than the H_m configuration on the listed black-box targets. These measurements concern the table's implementations, not a universal complexity separation or the runtime of the complete four-probe TCRA algorithm. Interpreting H_m requires distinguishing explicit Hessian construction from an HVP implementation.

PGN timing. With Inc-v3, batch size 10, and the reported RTX 2080 Ti setup, the measured times are 29.85 seconds for PGN and 30.95 seconds for TCRA, approximately 3.69% higher for TCRA. This observation is specific to the measured implementations; matching outer and inner iteration counts alone does not establish a matched gradient budget. Timing and gradient counts capture different aspects of cost.

H Relation to Gradient-Norm Regularization and PGN

PGN (Penalizing Gradient Norm) (Ge et al., 2023) uses local gradient-norm regularization to improve adversarial transfer. Gradient-norm regularization has also been studied for model generalization (Zhao et al., 2022). The use of finite differences or neighborhood samples alone is therefore not the distinctive component of TCRA. Its construction couples a response-guided second center, separation of parallel and transverse probe information, and loss-dependent aggregation across samples.

Proposition H.1 (Gradient-norm regularization as a separate objective)**.** *Let* $F(z) = J_s(z) - \lambda\|g(z)\|_2$*, with* $g(z) \neq 0$ *and* $v = -g(z)/\|g(z)\|_2$*. Under Lemma B.1,*

$$\left\|\nabla F(z) - \left[g(z) + \frac{\lambda}{s}\{g(z+sv) - g(z)\}\right]\right\|_2 \leq \tfrac{1}{2}\lambda\rho s. \tag{39}$$

Proof. Symmetry of the Hessian gives $\nabla F(z) = g(z) - \lambda H(z)g(z)/\|g(z)\|_2 = g(z) + \lambda H(z)v$. Lemma B.1 and $\|v\|_2 = 1$ yield the claimed error bound. □

The finite-difference expression is $(1-\lambda/s)g(z)+(\lambda/s)g(z+sv)$. It is a convex combination only if $0 \leq \lambda/s \leq 1$. This proposition concerns L_2-normalized probes and a gradient-norm penalty, not the L_1-normalized four-probe algorithm or the penalty $\mathcal{C}_R$. In particular, TCRA's mixing parameter δ is not identified with λ/s through this separate derivation.

The role of tilting is also specific: it prioritizes larger sampled surrogate losses, while the probe construction determines each direction. It does not guarantee flatness or suppress model-specific directions by itself. Model-shared high-loss stability is a mechanism to evaluate across models, rather than a consequence of the weighting formula.

I Additional Experimental Results

I.1 Transformer Targets

Adversarial examples generated on Inc-v3 are transferred to ViT (Dosovitskiy et al., 2021), PiT (Heo et al., 2021), Visformer (Chen et al., 2021), and Swin (Liu et al., 2021). Table 8 reports a mean ASR of 62.60% for TCRA and 56.98% for PGN, a gain of approximately 5.62 percentage points. TCRA improves all four listed target entries. This supports transfer across the evaluated architecture families, without isolating the geometric mechanism responsible for the improvement.

Table 8: Untargeted attack success rate (%) of adversarial examples in Transformer-based models. The adversarial examples are generated on Inc-v3.

Attack	ViT	PiT	Visformer	Swin	Avg.
MI	18.9	18.1	23.7	22.2	20.73
NI	20.0	19.9	25.5	25.5	22.73
VMI	28.4	34.0	41.1	40.8	36.08
EMI	26.0	27.3	36.9	36.7	31.73
RAP	35.2	41.8	50.6	50.2	44.45
PGN	44.7	53.2	65.5	64.5	56.98
TCRA	**50.5**	**59.1**	**70.7**	**70.1**	**62.60**

I.2 Defense Models

We generate adversarial examples using an equal-weight logit ensemble of Inc-v3, Inc-v4, and IncRes-v2. Table 9 contains nine defense configurations, including ComDefend (Jia et al., 2019), JPEG (Guo et al., 2018), NIPs-r3 (Naseer et al., 2020), R&P (Xie et al., 2018), HGD (Liao et al., 2018), and Bit-Red (Xu et al., 2018), together with the entries labeled FD, NPR, and RS.

The mean over the nine displayed targets is 82.01% for TCRA, compared with 77.08% for PGN. TCRA attains the highest displayed ASR on eight targets; NI has the highest value on RS. All averages below are recomputed from the displayed per-target values. These results measure transfer against the evaluated defenses and do not themselves establish shared flatness.

I.3 CIFAR-10

On CIFAR-10 (Krizhevsky et al., 2009), the reported settings use $\epsilon = 8/255$, $T = 10$, and an attack step $\eta = 1/255$. This dataset-specific attack step differs from ϵ/T. Adversarial examples are generated on VGG-16, ResNet-50, and DenseNet-121. Table 10 reports higher ASRs for TCRA than the listed baselines across the displayed source–target pairs. These results extend the empirical evaluation beyond the ImageNet-compatible dataset; they do not independently establish a causal link between flatness and transfer.

Table 9: Untargeted ASR (%) on nine defense configurations using an Inc-v3/Inc-v4/IncRes-v2 source ensemble. Avg. is the unweighted mean of the nine displayed entries. Bold denotes the highest value in each column.

Attack	ComDefend	JPEG	NIPs-r3	FD	R&P	HGD	Bit-Red	NPR	RS	Avg.
MI	54.9	49.5	29.9	51.7	22.2	24.8	23.8	36.5	30.3	35.96
NI	56.9	50.8	29.3	53.6	23.1	22.3	23.9	38.1	**80.7**	42.08
VMI	72.4	71.5	58.1	67.8	50.6	54.3	39.0	44.9	35.6	54.91
EMI	78.2	74.1	69.1	74.8	60.1	64.8	47.6	46.8	38.8	61.59
RAP	89.5	88.1	81.0	79.6	73.1	72.3	54.6	49.6	41.7	69.94
PGN	93.7	91.3	88.3	85.7	83.6	82.5	72.1	51.3	45.2	77.08
TCRA	**96.9**	**95.4**	**92.5**	**90.9**	**88.8**	**87.7**	**77.9**	**56.5**	51.5	**82.01**

Table 10: Untargeted attack success rates (%) on the CIFAR-10 dataset for the attack methods in the single model setting. The adversarial examples are crafted on VGG-16, ResNet-50 (Res-50), and DenseNet-121, respectively.

Attack	MobileNet	VGG-19	GoogLeNet	Inc-v3	DenseNet-121	DenseNet-169	Res-34	Res-50
MI	52.18	57.56	47.29	52.74	40.96	42.40	41.72	41.93
NI	56.13	61.36	49.19	54.87	37.61	39.74	38.74	38.90
VMI	66.14	68.05	60.89	65.63	55.62	57.21	55.35	56.46
EMI	70.69	74.36	66.78	70.56	59.98	63.04	60.47	61.83
RAP	77.98	78.43	73.41	77.86	68.30	69.74	65.48	66.27
PGN	85.97	86.73	82.82	85.59	72.48	74.66	71.62	72.95
TCRA	**90.89**	**91.59**	**87.98**	**90.75**	**77.88**	**80.12**	**76.93**	**77.95**

(a) Untargeted attack success rates (%) for the adversarial examples crafted on VGG-16.

Attack	MobileNet	VGG16	VGG19	GoogLeNet	Inc-v3	DenseNet-121	DenseNet-169	Res-34
MI	70.42	67.37	65.8	63.06	69.02	72.39	73.34	67.78
NI	71.97	65.57	63.76	63.28	69.13	71.03	72.78	65.02
VMI	77.60	74.27	73.26	71.11	75.99	76.80	77.68	73.54
EMI	80.11	78.66	77.43	76.34	78.12	79.68	80.12	77.24
RAP	86.92	84.24	83.46	80.68	81.75	83.54	84.98	81.36
PGN	90.88	88.68	88.07	85.79	89.53	89.93	90.91	87.19
TCRA	**95.75**	**93.34**	**93.02**	**88.98**	**94.79**	**94.97**	**96.12**	**92.45**

(b) Untargeted attack success rates (%) for the adversarial examples crafted on Res-50.

Attack	MobileNet	VGG16	VGG19	GoogLeNet	Inc-v3	DenseNet-169	Res-34	Res-50
MI	63.47	60.64	60.08	57.39	63.35	71.09	61.99	67.37
NI	66.86	61.80	60.85	60.30	66.54	75.92	61.57	69.25
VMI	71.28	68.49	68.01	65.40	70.62	75.51	68.42	72.50
EMI	74.36	75.66	73.54	70.41	75.23	78.94	73.64	77.45
RAP	79.98	80.22	78.68	76.39	80.55	84.25	78.65	80.03
PGN	86.73	85.12	84.66	81.82	86.26	88.35	83.30	86.21
TCRA	**91.73**	**90.55**	**89.89**	**86.72**	**91.39**	**93.39**	**88.71**	**91.95**

(c) Untargeted attack success rates (%) for the adversarial examples crafted on DenseNet-121.