Seeking Flat Minima over Diverse Surrogates for Improved Adversarial Transferability: A Theoretical Framework and Algorithmic Instantiation

Meixi Zheng, Kehan Wu, Yanbo Fan, Rui Huang, *Member, IEEE*, Baoyuan Wu†, *Senior Member, IEEE*

***Abstract*—The transfer-based black-box adversarial attack setting poses the challenge of crafting an adversarial example (AE) on known surrogate models that remain effective against unseen target models. Due to the practical importance of this task, numerous methods have been proposed to address this challenge. However, most previous methods are heuristically designed and intuitively justified, lacking a theoretical foundation. To bridge this gap, we derive a novel transferability bound that offers provable guarantees for adversarial transferability. Our theoretical analysis has the advantages of *(i)* deepening our understanding of previous methods by building a general attack framework and *(ii)* providing guidance for designing an effective attack algorithm. Our theoretical results demonstrate that optimizing AEs toward flat minima over the surrogate model set, while controlling the surrogate-target model shift measured by the adversarial model discrepancy, yields a comprehensive guarantee for AE transferability. The results further lead to a general transfer-based attack framework, within which we observe that previous methods consider only partial factors contributing to the transferability. Algorithmically, inspired by our theoretical results, we first elaborately construct the surrogate model set in which models exhibit diverse adversarial vulnerabilities with respect to AEs to narrow an instantiated adversarial model discrepancy. Then, a *model-Diversity-compatible Reverse Adversarial Perturbation* (DRAP) is generated to effectively promote the flatness of AEs over diverse surrogate models to improve transferability. Extensive experiments on NIPS2017 and CIFAR-10 datasets against various target models demonstrate the effectiveness of our proposed attack.**

***Index Terms*—Black-box adversarial attack, adversarial transferability, flatness, model discrepancy.**

I. Introduction

DEEP neural networks (DNNs) are vulnerable to adversarial examples (AEs), where attackers add imperceptible perturbations to benign examples but make a model produce erroneous predictions [1]–[4]. Under the black-box setting, attackers have no information regarding possible future target models, and the adversarial transferability matters since it allows the attackers to attack target models by alternatively generating AEs from the surrogate models. However, as the attacker can not access the information of the target model, a potentially unmatched surrogate model may lead to rather limited attack capability of the transferred AE against the target model.

Previous works [5] attributed the unsatisfactory transferability to the overfitting of AEs to the surrogate models. In turn, it is essential that such AEs be optimized using methods that ensure that crafted perturbations do in fact transfer beyond the surrogate models. A plethora of transfer-based black-box adversarial attack methods have been proposed [5]–[9]. There are rich advances in improving AEs' transferability from optimization, feature, input-transformation, and model perspectives. Despite the progress made, the adversarial transferability suffers from a lack of a general theoretical understanding. As a result, the literature relies heavily on empirical heuristics, without theoretical guarantees. *Can we build the theoretical foundation to deepen our understanding of transfer-based attacks?*

To tackle this problem, in this paper we present a novel theoretical analysis of transfer-based attacks towards generalizing previous works and explicitly guiding algorithm design by deriving a transferability bound. We start by formalizing the task of interest as crafting an AE that attacks successfully on the target model distribution. This idea consists of minimizing a *target adversarial risk*, which corresponds to the expected error of an AE over the target model distribution. We then decompose it into a *surrogate adversarial risk* and a *transferability gap*. The surrogate adversarial risk measures the expected error over the surrogate model distribution and could be upper bound estimated by its empirical version and the loss landscape sharpness at the AE (cf. Theorem 3). The transferability gap accounts for the discrepancy between surrogate and target model distributions and could be upper bounded in terms of a novel discrepancy, the adversarial model discrepancy, which is based on a variational representation that lower bounds ϕ-divergences [10] and is tailored to capture "adversarially significant" distribution differences (cf. Theorem 2). Combining the two bounds, we derive a transferability bound on target adversarial risk which provides a theoretical guarantee on the adversarial transferability (cf. Theorem 4). This bound further implies that the adversarial transferability of AEs has a positive correlation with three key factors simultaneously: (1) the white-box attack performance of the AE, (2) the regularization for surrogate models, and (3) the ϕ-divergences between the surrogate and target model distributions, resulting in an attack framework generalizing previously popular attacks as special cases (cf. Equation 26). By comparing our bound with these methods through the lens of this framework, we find they typically control only one or two key factors of this framework, neither

Meixi Zheng and Baoyuan Wu are with School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China. Kehan Wu and Rui Huang are with School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China. Yanbo Fan is with Nanjing University.

email: meixizheng1@link.cuhk.edu.cn, kehanwu1@link.cuhk.edu.cn, fanyanbo0124@gmail.com, ruihuang@cuhk.edu.cn, wubaoyuan@cuhk.edu.cn.

†Corresponding author: Baoyuan Wu (wubaoyuan@cuhk.edu.cn).

of which is desirable nor sufficient to achieve satisfactory transferability. However, our bound considers the surrogate adversarial risk and transferability gap jointly and properly, providing a tighter and more comprehensive guarantee on the transferability of AEs.

From an algorithmic perspective, by instantiating the derived bound using the total variation distance and constructing the surrogate model distribution with multiple distributional components, we further demonstrate that transferability can be expected if one seeks a flat minimum of empirical surrogate adversarial risk, where the surrogate components are designed to have similar model behaviors regarding adversarial vulnerability to AEs, on average, to the future target model distribution. Inspired by our theoretical result, we design a novel transfer-based adversarial attack. In particular, we first diversify the adversarial vulnerabilities in surrogate models by accounting for both between-distribution diversity and within-distribution diversity to approximate those of future target models, thus controlling the surrogate-target shift. We then propose to inject a model-**D**iversity-compatible **R**everse **A**dversarial **P**erturbation (DRAP) into the attack procedure to effectively optimize the loss landscape flatness at the AE over a set of diverse surrogate models. We conduct extensive experiments to evaluate DRAP on NIPS2017 and CIFAR-10 datasets, covering untargeted and targeted attacks against both standard and adversarially trained models and show that (1) compared with 14 state-of-the-art baseline attacks, DRAP achieves significant improvements in attack success rates; (2) DRAP is scalable to be combined with previous methods to further boost transferability; (3) Both optimization signals, seeking flat minima and improving diversity, in DRAP contribute to the transferability, corroborating our theoretical findings.

Contributions This work is an extension of our previous conference paper [11], compared to which the most significant updates and contributions are three-fold:

- Theoretically, we prove that the difference between target adversarial risk and empirical surrogate adversarial risk is upper bounded by a sharpness penalty and a model discrepancy penalty. This result provides a theoretical foundation for the assumed relationship between flatness and transferability in RAP and further points out that considering loss landscape flatness and model diversity in adversarial vulnerability simultaneously is exactly when this paper will bring the original RAP from the lab to the real world.
- Algorithmically, we propose a theory-guided attack strategy DRAP as a correction of RAP. It generates reverse adversarial perturbations tailored to each of the diverse surrogate models, which are selected based on two dimensions of diversity, to effectively find a flat local minimum among them. The code is publicly available[1].
- Empirically, we demonstrate the soundness of our attack by conducting comprehensive experiments on NIPS2017 and CIFAR-10 datasets against various target models. We also perform ablative studies to further understand the contribution of the two optimization signals and to verify the relevance of our theoretical findings.

[1]https://github.com/SCLBD/blackboxbench

II. Related work

Transfer-based attacks are motivated by the observation that AEs generated to deceive the surrogate model can also deceive the target model, even when their architectures differ significantly, as long as both models are solving the same task [12]. One of the seminal work, Iterative Fast Gradient Sign Method (I-FGSM) [13], generates adversarial examples by iteratively performing the fast gradient step, establishing a solid foundation for this area of research. However, it has been shown that I-FGSM often converges to poor local minima, resulting in low transferability [6].

Optimization-based attacks To improve transferability, better optimization algorithms are proposed to escape from poor local minima and yield AEs with better transferability, such as MI-FGSM [7], NI-FGSM [5] and PI-FGSM [14]. Recently, the connection between the loss landscape flatness and transferability has been extensively studied empirically [9], [11], [15]. Unfortunately, few works build a clear theoretical relationship between them. Our previous work, RAP [11] is the seminal work pursuing flatness of loss landscape for AEs. This idea is further formulated as a min-max bi-level optimization problem. PGN [15] also intuitively assumes that AEs at flat local regions tend to have better transferability and penalizes the gradient norm. CWA [9] derives an optimization objective involving the minimization of Hessian matrix's F-norm, thus pursues the flatness to boost transferability through a SAM-like strategy [16].

Feature-based attacks Methods from this perspective distort intermediate layer features by designing a new loss function. ILA [17] aims to use the suboptimal perturbation found by a basic attack as a proxy, deviating from it to increase the perturbation norm. Since increasing the norm in the image space is perceptible, ILA opts to increase the norm in the feature space instead. FIA [18] generates AEs by distorting object-related features, where the feature importance is defined by gradient. Beyond FIA, NAA [19] provides more accurate measures of neuron importance.

Input-transformation-based attacks Relatedly, a wide range of methods aim to simulate diverse models by applying input transformations on benign images, thus mitigating overfitting to surrogate models. For instance, DI-FGSM [6] applies random resizing and padding with a certain probability. SI-FGSM [5] enhances transferability by scaling. Admix [20] incorporates information from images in other classes by combining two images in a master-slave manner. TI-FGSM [21] utilizes translational shifts on the input image. SSA [22] generates diverse spectrum saliency maps to augment models, while SIA [23] applies local transformations across different regions of input to generate more diverse transformed images.

Model-based attacks Meanwhile, several methods have been proposed to enhance transferability from the model-centric perspective. One primary category focuses on model tuning. For instance, DRA [24] trains a score network to estimate ground-truth data score and use the estimated score

to update AE through Langevin dynamics. GhostNet [25] dynamically generates a vast number of ghost networks by applying erosion to specific intermediate structures of the base network. Bayesian attack [26] models the Bayesian posterior of the surrogate model, enabling an ensemble of infinitely many models. Another category emphasizes fusion strategies, which aim to reconcile gradients from multiple surrogate models to better capture intrinsic transfer information. Notable methods include Logit-ensemble [7], which attacks multiple models simultaneously by fusing their logit outputs and SVRE [27], which reduces gradient variance within ensemble.

III. Preliminaries

A. Transfer-Based Adversarial Attack

We first introduce the preliminaries about adversarial examples and specify a threat model under the naive transfer-based black-box setting. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^k$ be the original feature space and the label space. Let $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ be the set of possible image classifiers for a given task, where each $f(\boldsymbol{x}, \boldsymbol{w}) \in \mathcal{F}$ is a classifier mapping $\mathcal{X}$ to $\mathcal{Y}$, parameterized by $\boldsymbol{w} \in \mathcal{W}$.

Consider the general setting in a black-box adversarial attack against the target model $\mathcal{M}_{\mathcal{T}} \in \mathcal{F}$. For a benign image $\boldsymbol{x} \in \mathcal{X}$ and its ground truth label $y \in \mathcal{Y}$, the objective of the adversary is to find a perturbation $\boldsymbol{\xi} \in \mathbb{R}^d$, leading to an adversarial example, *i.e.*, $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\xi}$, such that $\mathcal{M}_{\mathcal{T}}(\hat{\boldsymbol{x}}) \neq y$ (untargeted attacks) or $\mathcal{M}_{\mathcal{T}}(\hat{\boldsymbol{x}}) = y_t$ (targeted attacks) with $y_t \in \mathcal{Y}\backslash\{y\}$. Besides, due to the imperceptible requirement, $\hat{\boldsymbol{x}}$ should be constructed within the neighborhood of an input image $\boldsymbol{x}$, *i.e.*, $\hat{\mathcal{X}}_{\boldsymbol{x},\gamma} = \{\hat{\boldsymbol{x}} : \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_\infty \leq \gamma\}$, dubbed γ-*norm ball*. For clarity, hereafter we denote it as $\hat{\mathcal{X}}$. $\gamma \geq 0$ is a pre-defined perturbation magnitude, and $\|\cdot\|_\infty$ denotes the L_∞-norm. Among all adversarial attack strategies, transfer-based attacks stem from the observation that adversarial samples crafted to deceive a white-box surrogate model set $\mathcal{M}_{\mathcal{S}} \subset \mathcal{F}$ have the capability to deceive a black-box target model $\mathcal{M}_{\mathcal{T}}$, provided that they are engaged in solving identical tasks. Generally, naive transfer-based attacks choose a single or a subset of arbitrary DNNs as surrogate models. Let ℓ be the adversarial loss function, one can seek the AE by solving the constrained optimization problem on $\mathcal{M}_{\mathcal{S}}$:

$$\arg\min_{\hat{\boldsymbol{x}}} \frac{1}{|\mathcal{M}_{\mathcal{S}}|} \sum_{f_i(\cdot,\boldsymbol{w}_i)\in\mathcal{M}_{\mathcal{S}}} \ell\left(f_i(\hat{\boldsymbol{x}}, \boldsymbol{w}_i), y\right), \text{ s.t. } \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_\infty \leq \gamma. \tag{1}$$

The above $\ell(\cdot,\cdot)$ is often instantiated as the negative cross-entropy function for untargeted attacks, while the cross-entropy function $w.r.t.$ the target label y_t for targeted attacks.

B. PAC-Bayes Bound

We then introduce the PAC model. We assume a distribution $\mathcal{D}$ from which the training instances $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are independently sampled to form a set $\mathcal{M}$, a prior distribution $\mathcal{P}$ on an arbitrary concept $c \in C$ which is independent of the training set $\mathcal{M}$, and a posterior distribution $\mathcal{Q}$ on c which depends on $\mathcal{M}$. Given any instance $\boldsymbol{x}$ and concept c, the loss function of $\boldsymbol{x}$ on c is given by $\ell(\boldsymbol{x}, c) \in [0, 1]$. We define risk $\ell(c)$ to be the expectation over sampling $\boldsymbol{x}$ of $\ell(\boldsymbol{x}, c)$, *i.e.*, $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(\boldsymbol{x}, c)]$, and empirical risk $\hat{\ell}(c)$ to be $\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{x}_i, c)$.

Theorem 1. ***(PAC-Bayes [28], [29])*** *For any prior distribution $\mathcal{P}$ on the concept c, $0 < \delta < 1$, with probability $1 - \delta$ over the draw of training set $\mathcal{M}$ with size $n \in \mathbb{N}$, for any distributions $\mathcal{Q}$ on c, the following bound holds:*

$$\mathbb{E}_{\mathcal{Q}}[\ell(c)] \leq \mathbb{E}_{\mathcal{Q}}[\hat{\ell}(c)] + \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log\frac{n}{\delta}}{2(n-1)}}. \tag{2}$$

The PAC-Bayes theorem could bound the generalization error between test loss $\ell(c)$ and training loss $\hat{\ell}(c)$ of distribution $\mathcal{Q}$ on the concept c that depends on the training set, in terms of the Kullback-Leibler (KL) divergence between $\mathcal{P}$ and $\mathcal{Q}$. In transfer-based adversarial attacks, it may be tempting to directly use the PAC-Bayes theorem to derive the transferability bound. However, one of the cornerstone assumptions underlying the PAC's success is that "test" samples should share the same distribution as "training" samples. However, the independent and identically distributed (i.i.d.) assumption does not generally hold in the black-box setting. For instance, consider using ResNet-50 as a surrogate model and ViT as a target model: although both are trained on the same dataset, they differ substantially in architecture and training strategies. These differences induce a surrogate–target distribution shift at the model level, violating the i.i.d. assumption required by standard PAC-Bayes analysis. This model-level non-i.i.d. discrepancy contributes directly to the transferability gap and complicates the theoretical analysis. Consequently, naïvely applying PAC-Bayes under the i.i.d. assumption risks producing bounds that are misleading in the black-box setting.

C. ϕ-divergence

In light of unseen target models, we reformulate another inducement of AEs' unsatisfactory transferability as the surrogate-target model shift. A successful AE should hopefully behave robustly under the shift. A key component in tackling the shift is to study the difference between surrogate and target models. In our work, we define a new discrepancy between surrogate and target model distributions based on the variational representation of ϕ-divergences. Here we review with the definition of ϕ-divergence and its variational representation.

Definition 1. ***(ϕ-divergence [30])*** *Consider two probability distributions μ and ν with μ absolutely continuous $w.r.t.$ ν. Assume both distributions are absolutely continuous $w.r.t.$ measure $d\boldsymbol{w}$, with densities p_μ and p_ν, respectively, on domain $\mathcal{W} \subset \mathbb{R}^{|\boldsymbol{w}|}$. Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex, lower semi-continuous function satisfying $\phi(1) = 0$. The ϕ-divergence D_ϕ is defined as:*

$$D_\phi(\mu\|\nu) = \int p_\nu(\boldsymbol{w})\phi\left(\frac{p_\mu(\boldsymbol{w})}{p_\nu(\boldsymbol{w})}\right) d\boldsymbol{w}. \tag{3}$$

ϕ-divergence measures the difference between two given probability distributions. A large class of popular statistical divergences could be recovered from ϕ-divergences as special cases of Equation (3). For example, $\phi(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x} - 1|$ recovers the total variation (TV) distance, *i.e.*, $\text{TV}(\mu\|\nu) = \frac{1}{2}\int |p_\mu(\boldsymbol{w}) - p_\nu(\boldsymbol{w})| d\boldsymbol{w}$. $\phi(\boldsymbol{x}) = (\boldsymbol{x} - 1)^2$ recovers the Pearson χ^2 divergence, *i.e.*, $\chi^2(\mu\|\nu) = \int \frac{(p_\mu(\boldsymbol{w}) - p_\nu(\boldsymbol{w}))^2}{p_\nu(\boldsymbol{w})} d\boldsymbol{w}$ [31]. Note that ϕ-divergence also has a variational representation formula which converts its calculation into an optimization

problem over a function space, offering a valuable mathematical view for the similarity between probability distributions [10], [32].

Lemma 1. ***(Variational formula of ϕ-divergences, Theorem 1 [10])*** *Let ϕ^* be the Fenchel conjugate function of ϕ, i.e., $\phi^*(t) = \sup_{x\in\ dom\ \phi}\{xt - \phi(x)\}$). With G encompassing all bounded measurable functions, let $\mathcal{G}$ be the family of functions with*

$$G \subset \mathcal{G} \subset L^1(\mu). \tag{4}$$

For any family of transformations

$$\mathcal{T} \subset \left\{T = T(g),\ \textit{such that}\ T : \mathcal{G} \mapsto L^1(\mu)\right\}. \tag{5}$$

Then any ϕ-divergence can be written as:

$$\begin{aligned} D_\phi(\mu\|\nu) = \sup_{g\in\mathcal{G}}\{\sup_{T\in\mathcal{T}}\{&\mathbb{E}_{\boldsymbol{w}\sim\mu}[T\left(g(\boldsymbol{w})\right)] \\ &- \mathbb{E}_{\boldsymbol{w}\sim\nu}\left[\phi^*\left(T\left(g(\boldsymbol{w})\right)\right)\right]\}\}. \end{aligned} \tag{6}$$

Taking the affine transformation as an example, $T_{\alpha,t} = tg+\alpha$ with $\alpha, t \in \mathbb{R}$, leads to the variational formula:

$$\begin{aligned} D_\phi(\mu\|\nu) = &\sup_{g\in\mathcal{G},t\in\mathbb{R}} \mathbb{E}_{\boldsymbol{w}\sim\mu}[tg(\boldsymbol{w})] \\ &- \inf_{\alpha\in\mathbb{R}}\left\{\mathbb{E}_{\boldsymbol{w}\sim\nu}\left[\phi^*(tg(\boldsymbol{w})+\alpha)\right] - \alpha\right\}. \end{aligned} \tag{7}$$

The variational representation in Lemma 1 yields a lower bound of ϕ-divergence when $\mathcal{G}$ and $\mathcal{T}$ contain only a subset of all possible functions.

IV. A Theoretical Guarantee on Adversarial Transferability

In this section, we warm up by formalizing the transfer-based attack as a target adversarial risk minimization problem (Section IV-A). Through decomposing the target risk into the surrogate adversarial risk and the transferability gap, and deriving the bounds for each part (Section IV-B and IV-C), we derive a transferability bound that provides a theoretical guarantee on the adversarial transferability (Section IV-D). Finally, we establish an attack framework from our bound that generalizes previous works as special cases (Section IV-E). In the following, we mainly focus on discussing the interpretations and implications of the theorems, and we refer readers to *Appendix* A for proof details.

A. Formalizing Transfer-Based Attacks

We start by defining the model distributions and the notion of risks we are concerned with.

Definition 2. ***(Model distribution)*** *Let $\mathcal{F}$ be the set of possible model architectures for a given task, each function $\hat{f} \in \mathcal{F}$ is a parametric family of models, where $\hat{f}(\cdot, \hat{\boldsymbol{w}}) : \mathcal{X} \to \mathcal{Y}$ is an example with parameter $\hat{\boldsymbol{w}} \in \mathbb{R}^{|\hat{\boldsymbol{w}}|}$. The parameter space induced by $\hat{f}$ is $\hat{\mathcal{W}} = \{\hat{\boldsymbol{w}} : \hat{\boldsymbol{w}} \in \mathbb{R}^{|\hat{\boldsymbol{w}}|}\}$. We define a model distribution by a distribution over function $P(\hat{f}(\cdot, \hat{\boldsymbol{w}}))$, induced by a generic distribution $P(\hat{\boldsymbol{w}})$ over parameters $\hat{\boldsymbol{w}}$ combined with a model architecture $\hat{f}(\cdot, \hat{\boldsymbol{w}})$. Typically, different model distributions may have different architectures. We assume that there exists a function f with parameter space $\mathcal{W} \subseteq \mathbb{R}^{|\boldsymbol{w}|}$ so that arbitrary $P(\hat{f}(\cdot, \hat{\boldsymbol{w}}))$ can fit into its architecture with a converted parameter distribution $P(\boldsymbol{w})$, i.e., $P(f(\cdot, \boldsymbol{w})) = P(\hat{f}(\cdot, \hat{\boldsymbol{w}}))$. By doing so, we remark that $P(f(\cdot, \boldsymbol{w}))$ is sufficiently general so as to define any model distribution on a common space $\mathcal{W}$, such that any model distribution is absolutely continuous $w.r.t.$ measure $d\boldsymbol{w}$, with density function $p(\boldsymbol{w})$. For the sake of clarity, we hereinafter omit the function form f from $P(f(\cdot, \boldsymbol{w}))$ and instead use the distribution on the underlying parametrization $P(\boldsymbol{w})$ to describe a model distribution.*

In the context of transfer-based attacks, once the attacker builds the surrogate model set, the *surrogate model distribution* $P_{\mathcal{S}}$ is observed, with density $p_{\mathcal{S}}$. Since attackers could customize surrogate models, $P_{\mathcal{S}}$ could be defined as a set of distributional components, $i.e.$, $P_{\mathcal{S}} = \{P_{\mathcal{S}_i}\}_{i=1}^{I}$, I is the total number of surrogate components owned by the attacker. In the next section, we will show that more surrogate components help to produce better attack performance. For clarity, and without loss of generality, in this section we consider $P_{\mathcal{S}}$ integrally. At test time, the attacker is facing any possible target models from the unobserved *target model distribution* $P_{\mathcal{T}}$, with density $p_{\mathcal{T}}$.

Definition 3. ***(Adversarial risk and empirical adversarial risk)*** *Consider a loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_0^+$. Let $P_{\mathcal{D}}$ be a model distribution. Assuming the AE $\hat{\boldsymbol{x}}$ as defined in Section III-A, we can define its adversarial risk on $P_{\mathcal{D}}$ by:*

$$R_{\mathcal{D}}(\hat{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{D}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y))]. \tag{8}$$

$R_{\mathcal{D}}(\hat{\boldsymbol{x}})$ characterizes the attack failure of an AE on $P_{\mathcal{D}}$. We sample K i.i.d. models $\{\boldsymbol{w}_i\}_{i=1}^{K} \sim P_{\mathcal{D}}$, forming a set $\mathcal{M}_{\mathcal{D}}$ of size K. Given $\mathcal{M}_{\mathcal{D}}$, we can define an empirical adversarial risk for $\hat{\boldsymbol{x}}$ by:

$$R_{\hat{\mathcal{D}}}(\hat{\boldsymbol{x}}) = \frac{1}{K}\sum_{\boldsymbol{w}_i\in\mathcal{M}_{\mathcal{D}}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_i\right), y\right). \tag{9}$$

The adversarial risk measures the expected attack error that an AE made according to the model distribution. For both adversarial risk and empirical adversarial risk, higher values indicate worse attack performance. We consider the 0-1 loss, $i.e.$, $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$. In untargeted attacks, $\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y) = \mathbb{I}(f(\hat{\boldsymbol{x}}, \boldsymbol{w}) \neq y)$, and in targeted attacks, $\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y) = \mathbb{I}(f(\hat{\boldsymbol{x}}, \boldsymbol{w}) = y_t)$, where $\mathbb{I}(\text{event}) = 1$ if the event happens, and 0 otherwise.

The task of transfer-based attacks is to find an AE $\hat{\boldsymbol{x}}$ that successfully attacks target models drawn from $P_{\mathcal{T}}$, $i.e.$, to minimize its attack failures. We formalize untargeted transfer-based attacks as a risk minimization problem under $P_{\mathcal{T}}$, $i.e.$, seeking a $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}$ that minimizes the *target adversarial risk* defined as follows:

$$\min_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}} R_{\mathcal{T}}(\hat{\boldsymbol{x}}), \tag{10}$$

$$\text{where } R_{\mathcal{T}}(\hat{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]. \tag{11}$$

The risk definition for targeted attacks is analogously obtained by substituting y with the target label y_t. By unifying targeted and untargeted attacks within a single risk minimization framework, we restrict our following analysis to the untargeted case without loss of generality, and the analysis for targeted attacks can be trivially recovered by adopting the targeted risk.

Under the black-box setting, no information about $P_{\mathcal{T}}$ is available during the attack, making it impossible to optimize $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$. In practice, attackers commonly resort to an alternative risk minimization, $i.e.$, minimizing the *surrogate adversarial risk* $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$, which is measured over the self-chosen surrogate

distribution $P_{\mathcal{S}}$, with the expectation of achieving good transferability. $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$ is defined as follows:

$$R_{\mathcal{S}}(\hat{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y))]. \tag{12}$$

However, the surrogate model distribution and the inaccessible target model distribution may differ significantly. As a result, the transfer gap between the target adversarial risk and the empirical surrogate adversarial risk becomes even worse due to this distribution shift, making the attack performance unsatisfactory.

Risk Decomposition To derive a bound on the transferability to the target model distribution of an AE optimized under the surrogate model distribution, we first decompose the target adversarial risk as follows:

$$R_{\mathcal{T}}(\hat{\boldsymbol{x}}) = \underbrace{R_{\mathcal{T}}(\hat{\boldsymbol{x}}) - R_{\mathcal{S}}(\hat{\boldsymbol{x}})}_{\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}})} + R_{\mathcal{S}}(\hat{\boldsymbol{x}}). \tag{13}$$

According to the above decomposition, it is clear that solely minimizing the surrogate adversarial risk using some attack strategies cannot guarantee the decreasing of the target adversarial risk. The transferability gap $\mathcal{E}_{\text{trans}}$, which captures the dissimilarity between the surrogate and target model distributions relevant to the context of adversarial transferability, should also be taken into account. Thus, in the following, we will derive an upper bound of the target adversarial risk through deriving the upper bounds of the transferability gap and the surrogate adversarial risk separately.

B. Model-Discrepancy-Based Bound on Transferability Gap

Equation (13) tells that the transferability gap $\mathcal{E}_{\text{trans}}$ depends on the discrepancy between $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$. Thus, we first define a model discrepancy tailored to comparing model distributions in the context of transfer-based adversarial attacks, which is crucial for deriving the subsequent bound on the transferability gap and consequently designing our attack strategy. Specifically, according to the variational formula of the ϕ-divergences (cf. Lemma 1), we introduce the adversarial model discrepancy $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$, as follows.

Definition 4. ***(Adversarial model discrepancy)*** *For any surrogate model distribution $P_{\mathcal{S}}$ and target model distribution $P_{\mathcal{T}}$, any γ-norm ball $\hat{\mathcal{X}}$ and any $r \geq 0$, the localized adversarial space $\hat{\mathcal{X}}_r$ is defined as:*

$$\hat{\mathcal{X}}_r = \left\{ \hat{\boldsymbol{x}} \in \hat{\mathcal{X}} \mid R_{\mathcal{S}}\left(\hat{\boldsymbol{x}}\right) \leq r \right\}. \tag{14}$$

Based on $\hat{\mathcal{X}}_r$, let $\hat{\mathcal{G}}_r$ be a set of measurable functions, i.e., $\hat{\mathcal{G}}_r = \{\ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) : \hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r\}$. We define the adversarial model discrepancy $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$ between $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ as:

$$\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) = \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t \in \mathbb{R}} \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y)] - \inf_{\alpha \in \mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y) + \alpha)] - \alpha\}. \tag{15}$$

Restricting $\mathcal{G}$ to the subset $\hat{\mathcal{G}}_r$, $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ discrepancy can be interpreted as a lower bound of a general class of ϕ-divergences $D_{\phi}(P_{\mathcal{T}} \| P_{\mathcal{S}})$, this property is crucial for deriving a general attack framework in Section IV-E. It's also easy to see that $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ is a monotonically increasing function $w.r.t.\ 0 \leq r \leq 1$. Furthermore, $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ has some properties. (1) Taking the supremum over $\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r$, this discrepancy restricts its attention to a localized adversarial space, within which the examples may commit low attack errors—an aspect of interest in the context of adversarial attacks. (2) $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) \geq 0$, the equality holds when $P_{\mathcal{S}} = P_{\mathcal{T}}$. To explicitly see this, we first consider $t = 0$. By Lemma 2, $\inf \phi^*(\alpha) - \alpha = 0$, leads to $\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y)] - \inf_{\alpha \in \mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y) + \alpha)] - \alpha\} = 0$ when $t = 0$, then we prove the non-negativity of $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$. Moreover, when $P_{\mathcal{S}} = P_{\mathcal{T}} = P$, $\mathbb{E}_{\boldsymbol{w}\sim P}[t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y) + \alpha] - \mathbb{E}_{\boldsymbol{w}\sim P}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y) + \alpha)] \leq 0$ by $\phi^*(\boldsymbol{x}) \geq \boldsymbol{x}$, leads to $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P \| P) = 0$.

We are now ready to provide a bound on the transferability gap $\mathcal{E}_{\text{trans}}\left(\hat{\boldsymbol{x}}\right)$ in terms of the proposed $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$ discrepancy.

Theorem 2. ***(Transferability gap bound)*** *Define $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) = \inf_{\alpha} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\phi^*\left(t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right]$. Given the surrogate model distribution $P_{\mathcal{S}}$ and target model distribution $P_{\mathcal{T}}$, for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant $c_1, c_2 \in [0, +\infty)$ subjected to the constraint $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(c_1) \leq c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y\right)\right]$, we have*

$$\mathcal{E}_{\textit{trans}}\left(\hat{\boldsymbol{x}}\right) \leq \frac{1}{c_1} \mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right) + c_2 r. \tag{16}$$

Furthermore, if $P_{\mathcal{S}}$ is a mixture distribution of I distributions, i.e., $P_{\mathcal{S}} = \frac{1}{I}\sum_{i\in[I]} P_{\mathcal{S}_i}$, then

$$\mathcal{E}_{\textit{trans}}\left(\hat{\boldsymbol{x}}\right) \leq \frac{1}{c_1 I} \sum_{i\in[I]} \mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}_i}\right) + c_2 r. \tag{17}$$

Theorem 2 bounds the transferability gap in terms of the adversarial model discrepancy between the surrogate and target model distributions, as well as a constant term related to localized adversarial space parameter r. To minimize $\mathcal{E}_{\text{trans}}$, one can use a small r, which aligns with the surrogate risk minimization that an attack strategy might aim to achieve in practice, thereby reducing both terms.

As an instantiation of Theorem 2, we consider the case of TV, namely $\mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$. We have the following result:

Corollary 1. *Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Given the surrogate model distribution $P_{\mathcal{S}}$ and target model distribution $P_{\mathcal{T}}$, for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant c_1 satisfying $0 \leq c_1 \leq 1$, we have*

$$\mathcal{E}_{\textit{trans}}\left(\hat{\boldsymbol{x}}\right) \leq \frac{1}{c_1} \mathrm{D}_{\textit{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}), \tag{18}$$

where $\mathrm{D}_{\textit{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) = \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r} |\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y)] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}', \boldsymbol{w}), y)]|$. (19)

According to risk decomposition, we have $R_{\mathcal{T}}(\hat{\boldsymbol{x}}) = \mathcal{E}_{\text{trans}} + R_{\mathcal{S}}(\hat{\boldsymbol{x}})$, we need to minimize $\mathcal{E}_{\text{trans}}$ and $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$ simultaneously to assure the target attack performance. Combining $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$ with the bound for $\mathcal{E}_{\text{trans}}$, we have the following bound on the target adversarial risk:

$$R_{\mathcal{T}}(\hat{\boldsymbol{x}}) \leq R_{\mathcal{S}}(\hat{\boldsymbol{x}}) + \frac{1}{c_1} \mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right) + c_2 r. \tag{20}$$

The above bound yields the following result: Let $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ be an AE optimized by minimizing risk on the surrogate mixture $P_{\mathcal{S}}$. If $\hat{\boldsymbol{x}}$ can successfully attack over $P_{\mathcal{S}}$ seen during optimization, then $\hat{\boldsymbol{x}}$ has bounded risk over future target model distribution $P_{\mathcal{T}}$, if $P_{\mathcal{T}}$ has low adversarial model discrepancy with $P_{\mathcal{S}}$.

C. PAC-Bayesian Bound on Surrogate Adversarial Risk

Taking advantage of the above result, we are now able to conduct a transferability-guaranteed attack. In practice, the attacker typically only owns a finite surrogate set $\mathcal{M}_\mathcal{S}$ of size K, *i.e.*, $\{\boldsymbol{w}_k\}_{k=1}^K \sim P_\mathcal{S}$, and the bound in Equation 20 needs to be estimated empirically. Hence, the next step in this section involves obtaining an empirical version of the bound in Theorem 20. Note that we present a concentration result solely for $R_\mathcal{S}(\hat{\boldsymbol{x}})$, as the target models are unavailable and the $\mathcal{E}_{\text{trans}}$-related terms are thus intractable. Introducing its computation offers no clear benefit. Nevertheless, these terms will serve to offer intuition for efficiently choosing surrogate models to control the transferability gap.

With the surrogate set $\mathcal{M}_\mathcal{S}$, minimizing *empirical surrogate adversarial risk* $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$,

$$R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}) = \frac{1}{K} \sum_{\boldsymbol{w}_k \in \mathcal{M}_\mathcal{S}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_k\right), y\right), \tag{21}$$

can have multiple local minima that provide similar white-box attack loss but significantly different generalization on $R_\mathcal{S}(\hat{\boldsymbol{x}})$, and consequently black-box performance $R_\mathcal{T}(\hat{\boldsymbol{x}})$. Unfortunately, the typical optimization methods, such as PGD [33] and I-FGSM [13], often lead to sub-optimal transferability [11].

To bound $R_\mathcal{S}(\hat{\boldsymbol{x}})$, our goal suggests that PAC-Bayes theorem in Theorem 1 may be fruitful. In our task, we have the concept is the adversarial example. The instance refers to the model. The risk refers to the adversarial risk. The generalization error measures how well the generated AE transfers from the employed samples to the surrogate model distribution (see these in Lemma 3). Under the PAC-Bayesian framework, we derive a bound for surrogate adversarial risk such that it could be estimated from finite models sampled from $P_\mathcal{S}$:

Theorem 3. ***(Surrogate risk bound)*** *For any $\rho > 0$, $0 < \delta < 1$, model distribution $P_\mathcal{S}$, and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$, with probability $1-\delta$ over the choice of surrogate model set $\mathcal{M}_\mathcal{S} \sim P_\mathcal{S}$ with size $K \in \mathbb{N}$, we have*

$$R_\mathcal{S}(\hat{\boldsymbol{x}}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) + \sqrt{\frac{\frac{d}{2}\log(1 + \frac{\gamma^2}{\rho^2}(1+\sqrt{\frac{\log K}{d}})^2) + \log\frac{K}{\delta} + \tilde{\mathcal{O}}(1)}{2(K-1)}}. \tag{22}$$

where $\tilde{\mathcal{O}}(1)$ term equals to $\varepsilon = \frac{1}{2} + 2\log(2 + 3d + 6r^2K + 4d\log(\sqrt{d} + \sqrt{logK}))$.

As we can see, this PAC-Bayes bound depends on two terms. The first one is the supremum of empirical surrogate risk over perturbed AE $\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}$, which denotes the worst-case attack error of neighborhood regions round $\hat{\boldsymbol{x}}$. The second one is the confidence bound which tells the effect of the number of surrogate samples K on transferability bound. If $\hat{\boldsymbol{x}}$ is optimized over enough samples, this term could be reduced, and one can use the first term as an upper bound estimator of surrogate risk.

D. Transferability Guarantees for Transfer-Based Attacks

Plugging the bounds in Theorem 2 and Theorem 3 into Equation 13 yields our final transferability PAC bound on target adversarial risk:

Theorem 4. ***(Transferability PAC bound)*** *Given the surrogate model distribution $P_\mathcal{S}$ and target model distribution $P_\mathcal{T}$. For any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant $c_1, c_2 \in [0, +\infty)$ subjected to the constraint $K_\mathcal{S}^{\hat{\boldsymbol{x}}}(c_1) \leq c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_\mathcal{S}}\left[\ell\left(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y\right)\right]$, with probability $1-\delta$ over surrogate model set $\mathcal{M}_\mathcal{S} = \{\boldsymbol{w}_j\}_{j=1}^K$ generated from distribution $P_\mathcal{S}$, we have*

$$R_\mathcal{T}(\hat{\boldsymbol{x}}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) + \frac{1}{c_1} \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_\mathcal{T} \| P_\mathcal{S}) + c_2 r + \varepsilon_{PAC}, \tag{23}$$

where $\varepsilon_{PAC} = \sqrt{\frac{\frac{d}{2}\log(1+\frac{\gamma^2}{\rho^2}(1+\sqrt{\frac{\log K}{d}})^2) + \log\frac{K}{\delta} + \tilde{\mathcal{O}}(1)}{2(K-1)}}$.

Rewriting the above bound, we have:

$$R_\mathcal{T}(\hat{\boldsymbol{x}}) \leq R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}) + \underbrace{\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) - R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})}_{\text{sharpness}} + \frac{1}{c_1} \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}\left(P_\mathcal{T} \| P_\mathcal{S}\right) + c_2 r + \varepsilon_{\text{PAC}}, \tag{24}$$

where the terms in the curly bracket depict the sharpness of $R_{\hat{\mathcal{S}}}$ at $\hat{\boldsymbol{x}}$ as it measures the difference of risk between $\hat{\boldsymbol{x}}$ and the worst-case point in the neighborhood of $\hat{\boldsymbol{x}}$ [34]. A low value of the sharpness term indicates that $\hat{\boldsymbol{x}}$ locates at the flat region of the loss landscape. As a result, the target adversarial risk of an AE $\hat{\boldsymbol{x}}$ can be bounded in terms of (1) the white-box attack performance of $\hat{\boldsymbol{x}}$ against $\mathcal{M}_\mathcal{S}$, (2) the sharpness of $R_{\hat{\mathcal{S}}}$ at $\hat{\boldsymbol{x}}$, (3) the adversarial model discrepancy, (4) a constant term related to localized adversarial space parameter r and (5) a confidence bound. Finally, we provide a guarantee that $\hat{\boldsymbol{x}}$ will "transfer well" on target model distribution $P_\mathcal{T}$, even when solely minimizing the empirical risk over the surrogate model set $\mathcal{M}_\mathcal{S}$. This bound inspires our basic plan of attack: controlling the adversarial model discrepancy, the attacker can conjure that finding a flat minimum on empirical surrogate adversarial risk will lead to better AE transferability.

One can also substitute Corollary 1 and Theorem 3 into Equation 13 to obtain the following concrete example of Theorem 4, specialized for $\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_\mathcal{T} \| P_\mathcal{S})$.

Corollary 2. *Given the surrogate model distribution $P_\mathcal{S}$ and target model distribution $P_\mathcal{T}$. For any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant c_1 satisfying $0 \leq c_1 \leq 1$, with high probability over surrogate model set $\mathcal{M}_\mathcal{S} = \{\boldsymbol{w}_j\}_{j=1}^K$ generated from distribution $P_\mathcal{S}$, we have*

$$R_\mathcal{T}(\hat{\boldsymbol{x}}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) + \frac{1}{c_1} \mathrm{D}_{TV}^{\hat{\mathcal{X}}_r}(P_\mathcal{T} \| P_\mathcal{S}) + \varepsilon_{PAC}. \tag{25}$$

E. A General Transfer-Based Attack Framework

In this section, we present an attack framework which generalizes previous transfer-based attacks. Through the lens of our framework, we revisit these attacks, especially RAP, and compare them with our transferability bound in Table I. The analysis shows that, while these attacks improve transferability through either finding better local minima in the surrogate loss landscape or tackling model shift, they do not consider both optimization signals simultaneously to achieve comprehensive transferability. Moreover, their principles of controlling the transferability gap are less tight than our $\mathrm{D}_\phi^{\hat{\mathcal{X}}_r}$ discrepancy and may result in unnecessary overestimation of the target risk bound. Experimental results confirm that considering surrogate

TABLE I
COMPARISON OF BOUNDS FOR TARGET ADVERSARIAL RISK $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$. THE "BOUND" MEANS THE RESPECTIVE ATTACK TRIES TO FIND AN AE $\hat{\boldsymbol{x}}$ WHICH MINIMIZES $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$ BY ALTERNATIVELY MINIMIZING THIS OBJECTIVE.

Method	Target risk $R_{\mathcal{T}}(\hat{\boldsymbol{x}}) \leq \underbrace{R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}) + \mathbf{r}(\mathcal{S})}_{\text{for bounding } R_{\mathcal{S}}(\hat{\boldsymbol{x}})} + \underbrace{\eta \mathrm{D}_{\phi}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)}_{\text{for bounding } \mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}})}$ +constant		
	Empirical surrogate risk $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$	Surrogate model regularization $\mathbf{r}(\mathcal{S})$	Transferability gap $\mathrm{D}_{\phi}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$
MI [7],NI [5],PI [14],VT [8]	Design elaborate optimizers to minimize $\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)$	\	\
CWA [9],SVRE [27]	Design elaborate optimizers to minimize $\frac{1}{\lvert\mathcal{M}_{\mathcal{S}}\rvert} \sum_{\boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_k\right), y\right)$	\	\
ILA [17]	$-\left(f_l\left(\hat{\boldsymbol{x}}'\right) - f_l(\boldsymbol{x})\right)\left(f_l\left(\hat{\boldsymbol{x}}\right) - f_l(\boldsymbol{x})\right)$	\	\
FIA [18]	$\sum\left(\bar{\Delta}_l^{\pi} \odot f_l\left(\hat{\boldsymbol{x}}\right)\right)$	\	\
NAA [19]	$\sum_{\substack{A_{l_j} \geq 0 \\ f_{l_j} \in f_l}} f_p\left(A_{l_j}\right) - \gamma \cdot \sum_{\substack{A_{l_j} < 0 \\ f_{l_j} \in f_l}} f_n\left(-A_{l_j}\right)$	\	\
PGN [15]	$\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)$	Sharpness $\lambda \cdot \max_{\lVert\boldsymbol{\epsilon}\rVert_p \leq \rho} \lVert\nabla_{\boldsymbol{\epsilon}} \ell\left(f\left(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}, \boldsymbol{w}\right), y\right)\rVert_2$	\
RAP [11]	$\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)$	Sharpness $\max_{\lVert\boldsymbol{\epsilon}\rVert_p \leq \rho} \ell\left(f\left(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}, \boldsymbol{w}\right), y\right) - \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)$	\
DI [6],TI [21],SI [5],Admix [20],SIA [23],SSA [22]	$\ell\left(f\left(\mathcal{T}(\hat{\boldsymbol{x}}), \boldsymbol{w}\right), y\right)$	\	Simulate different models with input transformations $\mathcal{T}$
GhostNet [25],Bayesian [26],LGV [35]	$\frac{1}{\lvert\mathcal{M}_{\mathcal{S}}\rvert} \sum_{\boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_k\right), y\right)$	\	Generating diverse variants from a base surrogate model
Our	$\frac{1}{\lvert\mathcal{M}_{\mathcal{S}}\rvert} \sum_{\boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_k\right), y\right)$	Sharpness $\max_{\lVert\boldsymbol{\epsilon}\rVert_p \leq \rho} \frac{1}{\lvert\mathcal{M}_{\mathcal{S}}\rvert} \sum_{\boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}, \boldsymbol{w}_k\right), y\right) - \frac{1}{\lvert\mathcal{M}_{\mathcal{S}}\rvert} \sum_{\boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}_k\right), y\right)$	$\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$

adversarial risk and transferability gap simultaneously and properly leads to significant gains (see Tables II, III).

Abstracted from our main result (cf. Equation 24), we give *a general framework* for bounding the target adversarial risk:

$$R_{\mathcal{T}}(\hat{\boldsymbol{x}}) \leq \underbrace{R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}) + \mathbf{r}(\mathcal{S})}_{\text{for bounding } R_{\mathcal{S}}(\hat{\boldsymbol{x}})} + \underbrace{\eta \mathrm{D}_{\phi}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)}_{\text{for bounding } \mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}})} + \text{constant}. \tag{26}$$

where η is a weight that trades off transferability with attack performance on surrogates. Within the first curly bracket, $\mathbf{r}(\mathcal{S})$ represents some form of regularization for surrogate models, $e.g.$, sharpness, which interacts with the empirical surrogate risk $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ to upper bound the surrogate risk $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$. Within the second curly bracket for bounding $\mathcal{E}_{\text{trans}}\left(\hat{\boldsymbol{x}}\right)$, we replace $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$ in Equation 24 by $\mathrm{D}_{\phi}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$ without violating the bound, as the variational representation $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$ is a lower bound of the ϕ-divergence D_{ϕ}.

Taking a second look at previous attacks within the above framework, some methods (such as MI, NI, PI, VT, CWA, SVRE) bound $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$ solely by $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ with one or several arbitrarily selected neural networks. Their introduced various gradient-based optimization algorithms to minimize $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ could help escape poor minima, thus improving $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$. Identically, methods from the feature perspective (such as ILA, FIA, NAA) focus on optimizing $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ by designing different loss functions to distort intermediate layer features rather than the final outputs. Going beyond simply accounting for the empirical risk, PGN and RAP bound $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$ by $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ in conjunction with specific surrogate model regularizations. However, the above methods falsely rely on an invalid i.i.d. assumption and overlook the transferability gap $\mathcal{E}_{\text{trans}}\left(\hat{\boldsymbol{x}}\right)$ in the target risk bound, thus achieving suboptimal attack performance.

Rather than relying solely on a few surrogate models, methods from input-transformation perspective (such as DI, TI, SI, Admix, SIA, SSA) and model perspective (such as GhostNet, Bayesian attack, LGV) augment a base surrogate model into an infinitely large set of models, aiming to align with those seen during inference. The common underlying assumption of these attacks is that the transferability of adversarial examples can be improved by attacking *more* models simultaneously, so efforts focus on obtaining as many different surrogate models as possible with a low computational cost [5], [23], [26]. By trying to place point masses at locations given by samples from the target model distribution, we can view these methods as simulating $P_{\mathcal{T}}$ with $P_{\mathcal{S}}$, which is in line with minimizing $\mathrm{D}_{\phi}\left(P_{\mathcal{T}} \| P_{\mathcal{S}}\right)$ in our framework, given that ϕ-divergences measure the difference between two given probability distributions. However, we demonstrate that D_{ϕ} may overestimate the transferability gap compared to our adversarial model discrepancy $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$, which only captures "practically significant" distribution difference. Moreover, despite mitigating the surrogate-target shift, these methods have not explored the AE optimizer in depth and typically default to the empirical risk minimization (ERM) approach I-FGSM.

In summary, existing attacks typically control only one of two relevant terms in Equation 26, neither of which is desirable nor sufficient to achieve satisfactory transferability. In contrast, our proposed attack accounts for $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$ and $\mathcal{E}_{\text{trans}}$ jointly and properly, providing a more tight and comprehensive guarantee on the transferability of AEs.

Detailed comparison with RAP [11] The relationship between flatness and transferability is initially explored in our prior work, RAP. However, this relationship is only intuitively assumed through an illustration (see Figure 1(b) in the RAP paper) and empirically validated without a rigorous theoretical grounding. In this work, our newly derived bound in Equation 24 provides a theoretical interpretation of the relationship between flat regions with low loss in loss landscape and the transferability of the AE, offering a solid theoretical assurance to the assumption of RAP. Moreover, inspired by this theoretical insight, our proposed attack plan in this paper advances RAP from two perspectives. First, we highlight the importance of the marginalization of surrogate models and seek flat minima over enough samples from the model distribution. Second, we explicitly require that surrogate models be carefully selected

to narrow the adversarial model discrepancy as small as possible, as sharpness by itself is insufficient to guarantee high transferability. From this point of view, RAP could be considered as a degraded version of our new attack.

V. Attack Algorithm

A. From the Bound to an Attack

We now exploit the above theoretical results to derive a novel practical attack algorithm. We will show how our theoretical analyses help us to make algorithmic extensions to the original RAP.

Inspired by Corollary 2, which instantiates Theorem 4 by a simple yet insightful TV distance, we first sample the surrogate model set from $P_{\mathcal{S}}$, where $P_{\mathcal{S}}$ is designed to minimize the TV adversarial model discrepancy with $P_{\mathcal{T}}$ such that $\mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}})$ is controlled. Then we propose to minimize the target adversarial risk $R_{\mathcal{T}}(\hat{\boldsymbol{x}})$ by the following optimization problem:

$$\min_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}} \max_{\|\boldsymbol{\epsilon}\|_p\leq\rho} \frac{1}{K} \sum_{\boldsymbol{w}_k\in\mathcal{M}_{\mathcal{S}}} \ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon},\boldsymbol{w}_k\right),y\right), \tag{27}$$

where ℓ is a surrogate loss to minimize the empirical surrogate adversarial risk. We generalize from L_2 norm to L_p norm to make it adaptive to the popular constraints ($i.e.$, L_2, L_∞) on examples in adversarial attacks. As a result, Equation 27 results in a strategy comprising two components: collecting surrogate models guided by the adversarial model discrepancy term and seeking a flat minimum $\hat{\boldsymbol{x}}$ according to a min-max term. We will discuss them separately in the following sections.

B. Narrow the Surrogate-Target Discrepancy

To make the attack power of an AE invariant across models, it is crucial that the surrogate model distribution $P_{\mathcal{S}}$ narrows the adversarial model discrepancy with $P_{\mathcal{T}}$. Specifically, the discrepancy term $\mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}})$ defined in Equation 19 quantifies the model discrepancy by identifying an input $\hat{\boldsymbol{x}}'$, which, being within $\hat{\mathcal{X}}_r$, implies its role as an adversarial example. This $\hat{\boldsymbol{x}}'$ attempts to differentiate between the expected adversarial loss $\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}',\boldsymbol{w}\right),y\right)\right]$ and $\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}',\boldsymbol{w}\right),y\right)\right]$, that is, separating $P_{\mathcal{S}}$ from $P_{\mathcal{T}}$ by comparing their error rates in robustness predictions. If no such $\hat{\boldsymbol{x}}'\in\hat{\mathcal{X}}_r$ exists, then we can consider $P_{\mathcal{S}}$ it as a sufficiently good approximation of $P_{\mathcal{T}}$. In this case, AEs generated by models from $P_{\mathcal{S}}$ are expected to exhibit strong transferability to models from $P_{\mathcal{T}}$. Ultimately, to efficiently control the transferability gap, the goal of $P_{\mathcal{S}}$ is to mimic the adversarial vulnerability of target models from $P_{\mathcal{T}}$, rather than faithfully approximate $P_{\mathcal{T}}$. The latter would require minimizing the TV distance, $\mathrm{D}_{\mathrm{TV}}(P_{\mathcal{T}}\|P_{\mathcal{S}}) = \int|p_{\mathcal{S}}(\boldsymbol{w})-p_{\mathcal{T}}(\boldsymbol{w})|d\boldsymbol{w}$, which provides a less tight bound to control the transferability gap as per in Section IV-E.

Between-distribution diversity We must carefully represent $P_{\mathcal{S}}$ in regions that contribute the most to mimicking the vulnerability of future target models in $P_{\mathcal{T}}$ $w.r.t.$ any $\hat{\boldsymbol{x}}'\in\hat{\mathcal{X}}_r$. To achieve this, we define $P_{\mathcal{S}}$ as a mixture of distributional components. Denote it as $P_{\mathcal{S}} = \frac{1}{I}\sum_{i\in[I]}P_{\mathcal{S}_i}$, where I is the total number of attacker-owned surrogate components. By applying the average-case transferability gap bound (cf. Equation 17 in Theorem 2) instantiated for $\mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_r}$ and following routine steps, we can easily yield an average-case transferability PAC bound by replacing $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}})$ in Theorem 4 with the averaged TV adversarial model discrepancy over multiple surrogate components, $i.e.$, $\frac{1}{I}\sum_{i\in[I]}\mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_r}\left(P_{\mathcal{T}}\|P_{\mathcal{S}_i}\right)$. This bound implies that: a flat minimum optimized over $\mathcal{M}_{\mathcal{S}}$ has a bounded risk $w.r.t.$ $P_{\mathcal{T}}$ that has similar model behaviors regarding adversarial vulnerability to AEs, *on average*, as the surrogate components $P_{\mathcal{S}_1},...,P_{\mathcal{S}_I}$. Naturally, this insight underscores *the importance of maximizing the diversity of the surrogate components* $\{P_{\mathcal{S}_i}\}_{i=1}^{I}$ $w.r.t.$ *adversarial vulnerability*, dubbed *between-distribution diversity*, since if two components exhibit similar vulnerability, one will be largely redundant in the averaging and contribute minimally to approximating the target vulnerability. Moreover, averaging over diverse components helps to smooth the risk from potentially unmatched surrogate choices.

Indeed, the adversarial vulnerability of DNNs is dominated by multiple factors, including model architectures and objective functions [36]. Attacks such as input-transformation-based approaches, which enhance the surrogate model space from a base model by applying transformations to its input, may introduce significant redundancy in model behaviors $w.r.t.$ adversarial vulnerability within surrogate models. In contrast, we prefer to combine multiple posterior distributions over the model weights, each independently trained on different architectures and with different training strategies, as components composing our surrogate model distribution. By doing so, we aim at enriching model behaviors in $P_{\mathcal{S}}$ $w.r.t.$ adversarial vulnerability. In particular, we must carefully choose surrogate components $\{P_{\mathcal{S}_i}\}_{i=1}^{I}$ so that their vulnerabilities could differ significantly and thus contribute efficiently to the overall goal. Empirical analyses have investigated the varying adversarial vulnerabilities across different models. From the perspective of architecture, [36], [37] indicates that the adversarial vulnerabilities of convolutional neural networks (convnets) and metaformers[2] differ significantly. From the perspective of training strategies, [36] observes that normally trained and adversarially trained models exhibit distinct types of adversarial vulnerabilities. These insights suggest that, from an adversarial standpoint, diversity in $P_{\mathcal{S}}$ can be efficiently obtained by simultaneously including the distributions of models from four prototypical categories (which we call *prototypes*): normal and adversarial versions of both convnet and metaformer. Model distributions across these prototypes exhibit significantly different vulnerabilities, making them ideal surrogate components.

Within-distribution diversity In practice, attackers need to sample surrogate models from each component to conduct their attacks, forming a finite surrogate set $\mathcal{M}_{\mathcal{S}}$ of size $K = In$, where n i.i.d surrogate models are sampled from each component $P_{\mathcal{S}_i}$, $i.e.$, $\left\{\boldsymbol{w}_j^i\right\}_{j=1}^{n}\sim P_{\mathcal{S}_i}$. Here, we further pursue the diversity of model samples from each component, dubbed *within-distribution diversity*. Despite operating within a single surrogate component, achieving within-distribution diversity remains critical. This ensures that each sampled points

[2]Metaformer [38] is a general architecture abstracted from Transformers and their variants.

contribute significantly to the approximation, maximizing the utility of each surrogate component. As observed in [39], the optima in the DNN loss surface are in fact connected, rather than isolated, forming a valley of low loss. This valley contains many high-performing and complementary models, which produce meaningfully different predictions, leading to a diverse variety of prediction behaviors. Therefore, samples from the distribution centered at the loss valley are preferred surrogate models. With constant learning rates, gathering the trajectory of weights traversed by SGD is approximately sampling from the distribution centered at the minimum of the loss [40]. Finally, we propose to gather SGD proposals during training models from each of the aforementioned four prototypes. By doing so, we achieve diversity within each surrogate component, while also maintaining diversity between components.

In a nutshell, considering within-distribution diversity and between-distribution diversity simultaneously is valuable in improving diversity in model behaviors $w.r.t.$ adversarial vulnerability to provide a better approximation to future target models.

C. Find a Flat Optimum from a Diverse Set of Surrogates

Even with the elaborately designed surrogate model set, simply crafting AEs with an ERM optimizer over these models can not be strong enough. Based on the perspective of loss landscape flatness in Theorem 4, we theoretically demonstrate that optimizing AE's flatness strengthens its transferability. An intuitive interpretation is that when pursuing a flat minimum among diverse models, it is more likely to remain in flat areas when applied to unseen target models. As a result, a small shift in the target model's loss landscape would not significantly increase the attack loss, making the AE less likely to fail. In this section, an optimization strategy, an upgraded version of the original RAP which is more compatible with a set of diverse surrogate models, is proposed to optimize flatness effectively and efficiently.

A general flatness-aware optimization Original RAP solves the bi-level optimization problem as in Equation 27,

$$\min_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}} \max_{\|\boldsymbol{\epsilon}\|_\infty\leq\rho} \frac{1}{K}\sum_{\boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}} \ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon},\boldsymbol{w}_k\right),y\right), \tag{28}$$

by iteratively optimizing the inner maximization and the outer minimization problem on the surrogate set $\mathcal{M}_\mathcal{S}$. In particular, at each iteration, fixing AE $\hat{\boldsymbol{x}}$, the inner maximization optimizes reverse perturbation $\boldsymbol{\epsilon}$ via a T-step I-FGSM. At each step, $\boldsymbol{\epsilon}$ is updated as follows:

$$\boldsymbol{\epsilon} \leftarrow \boldsymbol{\epsilon}+\beta_{\boldsymbol{\epsilon}}\cdot\text{sign}\left(\nabla_{\boldsymbol{\epsilon}}\frac{1}{K}\sum_{\boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}} \ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon},\boldsymbol{w}_k\right),y\right)\right), \tag{29}$$

where $|\mathcal{M}_\mathcal{S}|=K$, $\beta_{\boldsymbol{\epsilon}}$ is the inner step size and $\boldsymbol{\epsilon}$ is initialized by 0. Then, fixing reverse perturbation $\boldsymbol{\epsilon}$, the outer minimization update AE $\hat{\boldsymbol{x}}$ with the gradient calculated by minimizing the empirical surrogate adversarial risk $w.r.t.$ $\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}$:

$$\hat{\boldsymbol{x}} \leftarrow \Pi_\gamma\left[\hat{\boldsymbol{x}}-\beta_{\hat{\boldsymbol{x}}}\cdot\text{sign}\left(\nabla_{\hat{\boldsymbol{x}}}\frac{1}{K}\sum_{\boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}} \ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon},\boldsymbol{w}_k\right),y\right)\right)\right], \tag{30}$$

where $\beta_{\hat{\boldsymbol{x}}}$ is the outer step size, $\Pi_\gamma(\cdot)$ restricts current AE to be within a ℓ_∞-norm γ-ball of $\boldsymbol{x}$, and $\hat{\boldsymbol{x}}$ is initialized by the benign image. Note that after optimizing the loss of reversely perturbed AE $\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}$, we should come back to the center point $\hat{\boldsymbol{x}}$ to conduct this update.

Model-specific reverse perturbations In Section V-B, we enrich the surrogate model space to narrow the surrogate-target discrepancy. When optimizing the flatness over these diverse models, original RAP in practice computes a global reverse perturbation, $i.e.$, the $\boldsymbol{\epsilon}$ is maximized on an average of per-model losses and shared over the whole surrogate set $\mathcal{M}_\mathcal{S}$. However, with the diversity $w.r.t.$ adversarial vulnerability existed in $\mathcal{M}_\mathcal{S}$, each surrogate has its own worst-case reverse perturbation on AE and their optimization paths may conflict, directly optimizing a common reverse perturbation updated by fusing over a set of independent update directions will result in a weaker reverse perturbation than model-specific reverse perturbations. This motivates us to replace Equation 28 by calculating reverse perturbations of different models separately to improve the effectiveness of RAP:

$$\min_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}} \frac{1}{K}\sum_{\boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}} \max_{\|\boldsymbol{\epsilon}_k\|_\infty\leq\rho} \ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}_k,\boldsymbol{w}_k\right),y\right), \tag{31}$$

where $\boldsymbol{\epsilon}_k$ is calculated on individual models:

$$\boldsymbol{\epsilon}_k \leftarrow \boldsymbol{\epsilon}_k+\beta_{\boldsymbol{\epsilon}}\cdot\text{sign}\left(\nabla_{\boldsymbol{\epsilon}}\ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}_k,\boldsymbol{w}_k\right),y\right)\right), \boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}. \tag{32}$$

We call $\boldsymbol{\epsilon}_k$ a model-**D**iversity-compatible **R**everse **A**dversarial **P**erturbation (DRAP). Once DRAP is obtained, the outer minimization $w.r.t.$ $\hat{\boldsymbol{x}}$ in Equation 31 is performed. Given that $\mathcal{M}_\mathcal{S}$ may contain a large number of surrogate models (cf. Theorem 3), directly computing full gradients over the entire surrogate set at each iteration can be computationally inefficient. To address this, we adopt the longitudinal manner update [25], where $\hat{\boldsymbol{x}}$ is updated iteratively across models in $\mathcal{M}_\mathcal{S}$, one at a time. This reduces memory and computation overhead while maintaining the diversity-aware objective:

$$\hat{\boldsymbol{x}} \leftarrow \Pi_\gamma\left[\hat{\boldsymbol{x}}-\beta_{\hat{\boldsymbol{x}}}\cdot\text{sign}\left(\nabla_{\hat{\boldsymbol{x}}}\ell\left(f\left(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}_k,\boldsymbol{w}_k\right),y\right)\right)\right], \tag{33}$$

where each update step corresponds to a single surrogate model $\boldsymbol{w}_k\in\mathcal{M}_\mathcal{S}$. We alternate between generating $\boldsymbol{\epsilon}_k$ and updating $\hat{\boldsymbol{x}}$ across the model set, which effectively integrates model-specific reverse perturbations while ensuring scalability.

Improving optimizing stability In the first several iterations of generating the AE, solving the min-max problem in Equation 31 may hinder the AE efficiently converging to the region of high attack performance [11]. Evidence in Section VI-C1 shows that this phenomenon in RAP also exists in DRAP. A *late-start* strategy has been proposed by RAP that only solves outer minimization w.r.t unperturbed $\hat{\boldsymbol{x}}$ at the early stage, and then start RAP to seek flatness and further boost transferability. We also utilize this strategy during our optimization.

In addition to the late-start strategy, considering the diversity in models' loss landscapes inherent in DRAP, a velocity vector is accumulated in the gradient across iterations [41], with each iteration observing distinct surrogates, to stabilize the optimization path. Evidence soon presented in Section VI-C2 demonstrates that, although the idea of momentum is widely employed in previous attacks [5], [9], [14], our method could avoid the gradient overaccumulation which may hinder the attack [27], [36], [42] effectiveness and best benefits from it.

The complete pseudo-code of DRAP is shown in Algorithm 1.

Algorithm 1: Model-**D**iversity-Compatible **R**everse **A**dversarial **P**erturbation (DRAP) Algorithm

Require: benign data$(\boldsymbol{x}, y)$, perturbation budget γ, surrogate model distributions $\{P_{\mathcal{S}_i}\}_{i=1}^{I}$, number of samples within one component n, late start iteration number n_{LS}, inner step size $\beta_{\boldsymbol{\epsilon}}$, inner iteration number T, outer step size $\beta_{\hat{\boldsymbol{x}}}$, decay factor μ.
Initialize $\hat{\boldsymbol{x}} \leftarrow x, \boldsymbol{m} \leftarrow 0$;
for $j = 0, ..., n-1$ **do**
  for $i = 0, ..., I-1$ **do**
    Sample a surrogate model $\boldsymbol{w}_k$ from $P_{\mathcal{S}_i}$;
    if $j \geq n_{LS}$ **then**
      # Inner maximization
      Initialize $\boldsymbol{\epsilon}_k \leftarrow 0$;
      for $t = 0, ..., T-1$ **do**
        Update $\boldsymbol{\epsilon}_k$ using Equation 32;
      end for
    end if
    # Outer minimization
    Calculate $\boldsymbol{g} = \nabla_{\hat{\boldsymbol{x}}} \ell\left(f\left(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}_k, \boldsymbol{w}_k\right), y\right)$;
    Update momentum by $\boldsymbol{m} = \mu \cdot \boldsymbol{m} + \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_1}$;
    Update $\hat{\boldsymbol{x}} = \Pi_\gamma\left[\hat{\boldsymbol{x}} - \beta_{\hat{\boldsymbol{x}}} \cdot \operatorname{sign}(\boldsymbol{m})\right]$;
  end for
end for
return $\hat{\boldsymbol{x}}$.

VI. Experimental Evaluation

In this section, we conduct comprehensive evaluations to illustrate the soundness of DRAP. Specifically, our experiments are designed to explore the answers to the following questions:

1) *How does DRAP compare to previous ones when conducting untargeted and targeted attacks?*
2) *As a key property of transfer-based attacks, is DRAP scalable to be combined with input-transformation-based methods to further boost transferability?*
3) *Which aspect of DRAP's optimization bound, the sharpness penalty or model discrepancy penalty is the most important?*

We conduct the evaluations on ImageNet [43] and CIFAR-10 [44]. For ImageNet, we follow previous works [11], [15], [18], [21], [22], [25] and use the ImageNet-compatible dataset [3] in the NIPS 2017 adversarial competition, which contains 1,000 images with a resolution of $299 \times 299 \times 3$. For CIFAR-10, we conduct experiments on its test set with 10,000 images. In the following, we only consider ImageNet experiments. We put CIFAR-10 experiment results and its detailed experimental protocol in *Appendix* B.

A. Main Results

Baselines To answer the question 1), we take seven popular input-transformation-based attacks as our baselines, including I-FGSM [13], DI2-FGSM [6], SI-FGSM [5], Admix [20], TI-FGSM [21], SSA [22], SIA [23]. We also compare our attack with seven state-of-the-art optimization-based methods, namely MI-FGSM [7], PI-FGSM [14], VT-FGSM [8], RAP [11], PGN [15], CWA [9] and SVRE [27].

Models For surrogate models, we consider five architectures (*i.e.*, $I = 5$) from aforementioned four prototypical models: ResNet-50 [68] and ConvNeXt-T [54] from normally trained convnets, ViT [55] from normally trained metaformers, ResNet-50(AT) [69] from adversarially trained convnets and XCiT-S(AT) [67] from adversarially trained metaformers. ConvNeXt-T is included alongside ResNet-50 because it's a special convnet which follows designs popularized by vision transformers. For DRAP, model samples are gathered as proposals at each epoch during the fine-tuning of the five pretrained models, which are optimized using their respective training recipts over $n = 40$ additional epochs. In order to get more diverse samples, we fine-tune the five pretrained models with relatively larger constant learning rates, specifically 0.05, 0.001, 0.05, 0.5, and 0.001, respectively, for ResNet-50, ConvNeXt-T, ViT, ResNet-50(AT), and XCiT-S(AT), while without significantly degrading their clean accuracy. For compared methods, AEs are crafted on the five pretrained surrogate models by fusing the logits following [7]. To evaluate the transferability of AEs, we collect 31 target models to ensure comprehensive coverage of diverse model architectures from the four prototypical categories, abbreviated as ConvNet Set, Metaformer Set, ConvNet(AT) Set and Metaformer(AT) Set, as shown in Tab. II.

Implementation Details For the untargeted attack scenario, the adversarial perturbation is bounded by $\gamma = 4/255$ with step size $\beta_{\hat{\boldsymbol{x}}} = 2/255$ for all methods. For the targeted attack scenario, the adversarial perturbation is bounded by $\gamma = 16/255$ with step size $\beta_{\hat{\boldsymbol{x}}} = 8/255$ for all methods. We set the iteration number of MI, PI and CWA as 10 when conducting untargeted attacks, as suggested in their original papers, because their performance deteriorates for additional rounds. For RAP, the iteration number is 400. Otherwise, the iteration number is set as 200. For the hyper-parameters of DRAP, we set the number of samples within one model distribution $n = 40$, inner iteration number $T = 5$, late start iteration number $n_{LS} = 5$, inner step size $\beta_{\boldsymbol{\epsilon}} = 0.01/255$, decay factor $\mu = 1$. Note that the number of iterations for updating AE in DRAP is the same as others, as $n \times I = 200$. For compared methods, we follow the protocol in BlackboxBench benchmark.

Results of Untargeted Attacks We first summarize the untargeted attack results on ImageNet dataset against convnet set, metaformer set, adversarially trained convnet set and adversarially trained metaformer set, as shown in Table II. DRAP achieves a substantial improvement in the average attack success rate across all target models compared to the input-transformation-based methods and other optimization based methods. Taking a closer look at the comparison results, we found that, equipped with the same surrogate models, the state-of-the-art attack SIA is competitive on the relatively easier-to-attack normally trained target model sets. However, its performance is unsatisfactory when attacking the two adversarially trained target model sets than DRAP. DRAP provides a larger performance gain on attacking models with the defense mechanism while maintaining an acceptable performance on normal models, striking the balance among the whole target model sets. These results suggest that striving

[3] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/dataset

TABLE II
UNTARGETED ATTACK SUCCESS RATES (%,↑) ON IMAGENET DATASET. THE AES ARE CRAFTED FROM FIVE SURROGATE MODELS (RESNET-50, CONVNEXT-T, VIT, RESNET-50(AT)AND XCIT-S(AT)), AGAINST 31 TARGET MODELS FALLING INTO FOUR PROTOTYPES (NORMALLY AND ADVERSARIALLY TRAINED CONVNETS AND METAFORMERS). **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Target Model Set		I-FGSM	DI2-FGSM	SI-FGSM	Admix	TI-FGSM	SSA	SIA	MI-FGSM	PI-FGSM	VT-FGSM	PGN	CWA	SVRE	RAP	**DRAP**
ConvNet Set	AlexNet [45]	44.7	47.6	46.7	49.6	45.7	53.1	55.5	49.2	49.7	44.9	49.2	57.2	46.6	46.6	**68.5**
	VGG-16-BN [46]	52.7	66.4	66.3	81.1	57.4	75.5	**95.6**	68.1	71.8	58.3	81.7	66.7	57.9	54.4	84.1
	DenseNet-201 [47]	40.3	56.6	57.9	71.9	46.5	61.0	**90.8**	59.5	62.2	47.1	74.4	59.7	51.0	44.5	81.5
	GoogLeNet [48]	32.2	42.8	42.2	55.0	35.3	53.7	73.0	45.6	48.4	36.1	56.1	50.4	38.1	36.7	**73.9**
	ShuffleNetV2 [49]	42.7	51.9	52.4	63.9	44.2	62.2	77.3	54.8	56.7	45.1	63.1	61.2	48.6	46.1	**82.1**
	MobileNetV2 [50]	47.6	62.4	61.5	75.3	52.2	70.7	**93.6**	63.4	68.6	53.9	76.3	67.3	56.4	51.7	87.2
	MobileNetV3-L [51]	33.5	49.5	45.5	60.5	37.4	65.1	83.1	48.5	52.8	38.0	64.8	59.6	41.4	39.2	**85.5**
	MNASNet [52]	42.1	57.2	56.8	72.9	46.6	68.2	**92.3**	58.8	63.8	48.8	72.1	63.8	52.2	49.0	87.6
	EfficientNet [53]	31.5	46.8	40.7	49.6	35.0	56.2	**75.9**	44.7	46.6	35.1	55.8	52.7	38.4	36.8	74.9
	ConvNeXt-L [54]	36.3	50.2	45.7	66.5	36.8	68.3	**91.4**	58.0	59.1	42.7	67.8	57.6	50.9	46.7	77.4
	Average	40.4	53.1	51.6	64.6	43.7	63.4	**82.9**	51.7	54.5	45.0	66.1	59.6	48.2	45.2	80.3
Metaformer Set	ViT-S [55]	10.0	20.2	13.1	18.8	12.3	24.9	**40.2**	19.2	20.2	11.6	22.8	22.1	16.1	16.9	38.8
	DeiT-S [56]	14.1	26.5	17.3	23.5	17.3	32.4	45.2	25.2	25.1	15.6	26.9	29.4	20.1	20.0	**53.5**
	PoolFormer-S [38]	29.0	49.6	37.2	51.5	33.4	62.1	**86.2**	46.0	49.9	33.4	58.1	46.6	40.3	36.2	71.8
	TNT-S [57]	13.3	26.8	17.6	25.5	16.2	36.1	**57.3**	25.5	26.2	16.5	29.3	27.8	21.5	22.2	52.5
	Swin-S [58]	8.8	19.8	11.9	17.9	11.0	26.5	**42.9**	18.1	17.6	11.0	20.7	18.4	14.0	15.4	28.5
	XCiT-S [59]	11.3	27.9	13.2	16.1	11.8	29.1	**43.0**	18.4	18.9	12.2	19.7	20.3	16.5	16.3	30.6
	CaiT-S [60]	5.0	19.2	6.8	8.6	6.1	17.4	**29.4**	10.1	10.4	6.2	12.2	12.5	7.8	9.5	22.8
	Average	13.1	27.1	16.7	23.1	15.4	32.6	**49.2**	24.1	25.1	15.2	27.1	25.3	19.5	19.5	42.6
ConvNet(AT) Set	RaWideResNet-101-2 [61]	17.0	17.6	17.3	17.4	17.1	19.5	17.7	19.0	18.7	17.1	17.8	24.2	17.8	16.3	**26.8**
	WideResNet-50-2 [62]	21.9	22.6	22.1	22.1	22.3	25.3	23.1	24.0	24.0	22.0	23.5	32.5	23.3	22.0	**34.3**
	ResNet-50 [63]	39.7	40.2	39.9	40.4	40.6	43.4	41.9	41.7	41.8	39.6	41.3	47.6	41.1	39.7	**51.4**
	ConvNeXt-L [64]	10.4	10.8	10.5	10.8	10.8	12.5	11.3	11.7	11.7	10.4	11.5	16.4	11.2	10.7	**17.6**
	ConvNeXt-B [64]	10.6	10.9	10.4	11.0	11.1	13.3	11.2	12.3	12.4	10.7	12.0	17.8	11.3	11.1	**18.8**
	ConvNeXt-L-ConvStem [65]	10.2	10.6	9.9	10.7	10.3	12.5	10.8	11.2	11.1	10.1	11.0	15.9	10.7	10.5	**16.8**
	ConvNeXt-B-ConvStem [65]	11.7	11.8	11.7	12.0	12.2	14.4	12.8	12.9	12.8	11.4	12.4	18.7	12.2	11.2	**19.8**
	Inc-v3$_{ens3}$ [66]	9.7	13.7	10.3	10.9	9.8	19.1	17.8	12.7	13.8	10.1	13.0	17.1	11.6	11.6	**24.2**
	Inc-v3$_{ens4}$ [66]	11.4	15.7	12.8	14.0	11.6	21.5	20.0	14.8	14.9	11.0	14.9	19.0	12.0	13.3	**26.4**
	IncRes-v2$_{ens}$ [66]	3.6	6.4	4.6	5.5	3.9	10.4	8.7	6.1	6.5	4.0	7.3	9.0	4.5	5.6	**13.2**
	Average	14.6	16.0	15.0	15.5	15.0	19.2	17.5	16.3	16.4	14.6	16.5	21.8	15.6	15.2	**24.9**
Metaformer(AT) Set	Swin-B [64]	11.5	12.3	11.9	11.7	12.0	13.8	12.1	12.8	12.8	11.7	12.4	17.3	12.5	11.9	**18.6**
	Swin-L [64]	9.8	10.0	10.0	9.9	9.9	11.5	10.8	10.6	10.5	10.0	10.7	14.8	10.7	10.1	**15.6**
	XCiT-L [67]	14.9	15.5	15.0	15.3	15.2	18.4	16.0	17.2	17.2	14.9	17.2	26.8	16.0	14.4	**28.0**
	ViT-B-ConvStem [65]	11.2	11.6	11.5	11.6	11.3	12.9	12.1	12.4	12.2	11.4	12.4	17.8	11.8	11.7	**18.6**
	Average	11.9	12.4	12.1	12.1	12.1	14.2	12.8	13.3	13.2	12.0	13.2	19.2	12.8	12.0	**20.2**
Overall Average		22.2	30.0	26.8	32.6	24.0	35.8	45.1	30.5	31.7	24.2	34.5	34.5	26.6	25.4	**46.2**

for flatness among all surrogate models meanwhile considering model diversity could provide a strong guarantee on the transferability of AEs, regardless of the robustness of the target model sets. CWA is also shown as a promising prior method which could effectively utilize all diverse surrogate models simultaneously by attacking their common weakness and optimizing the flatness, leading to improved attack performance on the challenging adversarially trained model sets. However, it fails to explicitly address the surrogate-target model gap. Furthermore, its use of a universal perturbation across the model ensemble may hinder the optimization of flatness, resulting in an overall lower attack success rate compared to DRAP. In Section VI-C1, we further explore various solvers for optimizing flatness among diverse models. To sum up, we view the consistency with which DRAP outperforms the best prior methods, which change across different model sets, as a major advantage of the proposed method. Notably, in this experiment, the unseen target models are drawn from the same prototypical model sets as the surrogate models. In the ablative study in Section VI-C1, we consider stricter attack scenarios where the surrogate-target model shifts are more prominent to further evaluate the effectiveness of DRAP.

Results of Targeted Attacks The targeted attack results of baseline attacks and DRAP are shown in Table III. At first glance, it is evident that targeted attacks pose a more challenging scenario, particularly on the two model sets equipped with defense mechanisms. Despite using a looser perturbation constraint, most methods fail to induce even a single misclassification as the target label on the defense models. Among the prior methods, SIA demonstrates competitive performance in the targeted setting; for instance, it achieves relatively larger improvements on the two normal model sets, with success rates of 62.0% and 41.5%, respectively. CWA is the only prior method that reports success on the challenging defense model sets, achieving success rates of 6.4% and 7.5%. DRAP, however, achieves the best attack success rate by a significant margin as it is more successful at attacking various target models under the targeted scenario. This indicates that DRAP is also more effective under the targeted attack scenario, highlighting its ability to synergize well with different attack targets.

B. Composition with Input-Transformation-Based Attacks

Instead of improving transferability from an optimization perspective as considered in DRAP, another related panoply of methods introduce randomness into input via various transformations. Prior research has demonstrated that combining these two perspectives could achieve state-of-the-art transferability. As the outer minimization with $\hat{\boldsymbol{x}}$ in DRAP, $i.e.$, the Equation 33, could be solved by any off-the-shelf strategies, including the input-transformation-based methods, our method could also be seamlessly combined with them. To answer question 2), we explore the behavior of DRAP, as well as some well-known optimization-based attacks such as MI-FGSM [7], PI-FGSM [14], VT [8], our original RAP [11], PGN [15] and CWA [9], when combined with input-transformation-based attacks, namely, DI-FGSM [6], TI-FGSM [21], Admix [20]

TABLE III
TARGETED ATTACK SUCCESS RATES (%,↑) ON IMAGENET DATASET. THE AES ARE CRAFTED FROM FIVE SURROGATE MODELS (RESNET-50, CONVNEXT-T, VIT, RESNET-50(AT)AND XCIT-S(AT)), AGAINST 31 TARGET MODELS FALLING INTO FOUR PROTOTYPES (NORMALLY AND ADVERSARIALLY TRAINED CONVNETS AND METAFORMERS). THE RESULTS ARE AVERAGED ON EACH MODEL SETS. FULL RESULTS BROKEN DOWN INTO EACH MODELS ARE SHOWN IN *Appendix* E. **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Target Model Set	I-FGSM	DI2-FGSM	SI-FGSM	Admix	TI-FGSM	SSA	SIA	MI-FGSM	PI-FGSM	VT-FGSM	PGN	CWA	SVRE	RAP	**DRAP**
ConvNet Set	5.4	22.0	11.6	11.3	8.2	29.8	62.0	5.3	9.7	6.9	37.4	36.0	18.5	14.5	**77.6**
Metaformer Set	0.5	11.6	1.7	1.6	1.1	16.5	41.5	1.0	1.8	1.1	13.3	21.9	10.5	3.1	**56.4**
ConvNet(AT) Set	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	0.1	0.0	0.1	6.4	0.1	0.1	**13.4**
Metaformer(AT) Set	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	7.5	0.0	0.2	**11.4**
Overall Average	1.8	9.7	4.1	4.0	2.9	13.4	29.4	1.9	3.6	2.5	15.1	19.5	8.4	5.4	**43.5**

and SIA [23]. We omit the combination of PGN and SIA due to the heavy computational demands. The experimental protocol follows the untargeted setting in Section VI-A. The resultant composite attack performance is shown in Table IV. As can be observed, for both convnets and metaformers and for both normally trained and adversarially trained models, combining DRAP with existing input-transformation-based attacks can significantly improve the base version, leading to a new state-of-the-art attack performance. For example, SIA achieves a competitive average attack success rate of 45.1% (cf. Table II), while integrating with DRAP further improves it by a clear margin of 6.0%. These remarkable improvements validate the scalability of our method when combined with others to further boost adversarial transferability. Additionally, we view the consistency with which the extensions of DRAP outperform those of prior optimization-based methods, confirming the superiority of DRAP.

C. Ablative Study

From the results above, we conclude that DRAP inspired from the theoretical bound could learn an AE with strong transferability toward target models. To answer question 3), we need to gain a deeper insight into the rationale behind its superior attack performance. In this subsection, we disentangle the two distinct optimization signals within the bound: a sharpness penalty for pursuing a flat local minimum $\ell_{sharp} = \max_{\|\boldsymbol{\epsilon}\|_\infty \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) - R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ and a model discrepancy penalty for narrowing the surrogate-target shift $\ell_{dis} = \frac{1}{I}\sum_{i\in[I]} \mathrm{D}_{\mathrm{TV}}^{\hat{\mathcal{X}}_T}(P_{\mathcal{T}} \| P_{\mathcal{S}_i})$. We conduct ablative studies to explore the impact of each aspect of DRAP: first to determine the importance of the sharpness penalty term in our bound and the effectiveness of proposed model-diversity-compatible optimization algorithm (cf. Algorithm 1), second to determine the importance of the model discrepancy penalty term in our bound and the effectiveness of the strategy to choose surrogate models. We evaluate these ablations on ImageNet, using the same untargeted experimental protocol as in Section VI-A.

1) On the Sharpness Penalty

Is the flatness beneficial for boosting the transferability? First we analyze the importance of optimizing flatness in boosting transferability. We formalize this study as ablating the sharpness penalty term ℓ_{sharp} from our optimization objective and evaluating the ablated objective by reporting the attack success rate of I-FGSM, DI-FGSM, TI-FGSM and Admix combined with DRAP. We use the same experimental protocol as in Section VI-A but with the ablated objective. The results are presented in Table V. Within each combination, the first row represents our combinational method with the full optimization objective, applying the sharpness penalty starting at iteration n_{LS}. The second row represents the same objective but applies the sharpness penalty from iteration 0, implying an ablation on the late-start strategy. The third row represents the combinational method without penalizing the sharpness of AE. Across all combinations, DRAP with complete objective consistently outperforms attacks that solely penalize model discrepancy, regardless of whether the late-start strategy is used. Furthermore, we see a stronger attack performance of DRAP with late start strategy, which helps stabilize convergence. The results validate our theoretical result from Theorem 4 that

TABLE IV
ATTACK SUCCESS RATES (%, ↑) OF MI, PI, VT, RAP, PGN, CWA AND DRAP, WHEN IT IS INTEGRATED WITH DI, TI, ADMIX AND SIA, RESPECTIVELY. THE INDENTATION DENOTES COMBINATION. THE RESULTS ARE AVERAGED ON EACH MODEL SETS.

Attack	ConvNet Set	MetaFormer Set	ConvNet (AT) Set	MetaFormer (AT) Set	*Overall Average*
DI-FGSM	53.1	27.1	16.0	12.4	30.0
+ MI	75.0	50.6	18.9	13.7	43.5
+ PI	67.1	40.0	18.5	13.7	38.4
+ VT	56.6	31.9	16.1	12.4	32.3
+ RAP	51.8	34.0	17.5	12.5	31.6
+ PGN	65.6	27.2	16.9	13.0	34.4
+ CWA	67.4	39.6	23.7	19.5	40.8
+ DRAP	**83.2**	**56.6**	**27.7**	**20.6**	**51.2**
TI-FGSM	43.7	15.4	15.0	12.1	24.0
+ MI	57.8	26.5	17.6	14.1	32.1
+ PI	60.2	27.4	17.6	13.9	33.1
+ VT	48.1	17.8	15.1	12.2	26.0
+ RAP	46.1	21.7	15.6	12.2	26.4
+ PGN	66.4	29.7	17.7	13.6	35.6
+ CWA	60.6	26.8	22.5	19.3	35.3
+ DRAP	**79.2**	**44.4**	**25.6**	**20.2**	**46.5**
Admix	64.6	23.1	15.5	12.1	32.6
+ MI	66.4	29.2	17.3	13.5	35.4
+ PI	64.0	27.1	17.2	13.4	34.0
+ VT	58.0	21.4	15.1	12.3	30.0
+ RAP	56.7	26.2	16.4	12.6	31.1
+ PGN	65.3	26.9	16.5	12.7	34.1
+ CWA	64.0	25.7	22.0	19.1	36.0
+ DRAP	**82.4**	**43.2**	**24.5**	**19.7**	**46.8**
SIA	82.9	49.2	17.5	12.8	45.1
+ MI	83.0	55.5	20.4	13.9	47.7
+ PI	82.8	52.0	20.1	13.8	46.7
+ VT	83.1	51.5	18.0	12.8	45.9
+ RAP	72.5	45.5	19.1	13.0	41.5
+ PGN	-	-	-	-	-
+ CWA	83.4	52.0	23.8	17.9	48.7
+ DRAP	**86.9**	**55.3**	**25.6**	**18.2**	**51.1**

TABLE V
ABLATING THE SHARPNESS PENALTY TERM (- ℓ_{sharp}) AND LATE START STRATEGY (- LATE START) FROM DRAP'S COMBINATIONS WITH I-FGSM, TI-FGSM, DI-FGSM AND ADMIX. THE RESULTS ARE AVERAGED ON EACH MODEL SETS. **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

DRAP	ConvNet Set	Metaformer Set	ConvNet (AT) Set	Metaformer (AT) Set	*Overall Average*
+ I-FGSM	**80.2**	**42.6**	**24.9**	20.2	**46.1**
- late start	78.8	42.2	**24.9**	**20.3**	45.6
- ℓ_{sharp}	77.7	36.3	24.1	20.0	43.6
+ TI-FGSM	**79.2**	**44.4**	25.6	**20.2**	**46.5**
- late start	78.1	43.2	**25.8**	**20.2**	45.9
- ℓ_{sharp}	77.8	38.5	24.8	20	44.3
+ DI-FGSM	**83.2**	**56.6**	27.7	20.6	**51.2**
- late start	81.3	55.0	**28.0**	**20.7**	50.3
- ℓ_{sharp}	82.7	53.2	27.1	20.5	50.1
+ Admix	**82.4**	**43.2**	**24.5**	**19.7**	**46.8**
- late start	81.4	40.5	**24.5**	19.6	45.8
- ℓ_{sharp}	81.7	40.1	24.0	**19.7**	45.7

controlling the surrogate-target shift, finding a flat minima of surrogate adversarial risk lead to improved transferability. A detailed parameter study on the impact of late start iteration number n_{LS} is provided in *Appendix* D-A.

How to effectively optimize flatness of AE across a diverse set of surrogate models? Given a set of diverse surrogate models obtained following Section V-B, aside from our proposed algorithm (cf. Algorithm 1), there are other possible solvers to optimize flatness across this set. Here, we consider two implementations inspired by RAP and CWA, both of which aim to generate adversarial examples within flat local regions. We use Flat-RAP and Flat-CWA as the shorthands for optimizing flatness across diverse surrogate models using strategies of RAP and CWA, respectively.

To elaborate, our proposed algorithm boosts the flatness via a min-max bi-level optimization framework. It finds the worst-case reverse perturbation specific to each surrogate model at the inner step (refer to Equation 32) and updates AE toward the point where added with the model-specific perturbation could minimize the attack loss on the one surrogate model at the outer step (refer to Equation 33). This algorithm is expected to seek out AEs whose entire neighborhoods have uniformly low empirical surrogate adversarial risk value, $i.e.$, AEs locating at the flat regions of each of diverse models. However, from the perspective of RAP, Flat-RAP applies the inner maximization and outer minimization on the whole surrogate model set, $i.e.$, a global reverse perturbation (refer to Equation 29) and global update direction (refer to Equation 30) are obtained. Furthermore, though the lens of CWA, Flat-CWA substitutes the outer step of Flat-RAP with successively performing updates using each surrogate model to pursue a common weakness of model ensemble:

$$\hat{\boldsymbol{x}} \leftarrow \Pi_\gamma \left[\hat{\boldsymbol{x}} - \beta_{\hat{\boldsymbol{x}}} \cdot \operatorname{sign}\left(\nabla_{\hat{\boldsymbol{x}}} \ell\left(f\left(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}, \boldsymbol{w}_k\right), y\right)\right)\right], \boldsymbol{w}_k \in \mathcal{M}_{\mathcal{S}}, \tag{34}$$

where $\boldsymbol{\epsilon}$ is a global reverse perturbation same as in Flat-RAP. We provide their pseudocodes and implementation details in *Appendix* C. Note that though our derivation of Flat-RAP and Flat-CWA define the objective over the entire model set, in practice, we compute the gradient per-batch.

TABLE VI
ATTACK SUCCESS RATES (%, ↑) OF FLAT-RAP, FLAT-CWA AND DRAP. THE RESULTS ARE AVERAGED ON EACH MODEL SETS. **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Attack	ConvNet Set	Metaformer Set	ConvNet (AT) Set	Metaformer (AT) Set	*Overall Average*
Flat-RAP	77.4	**47.0**	19.9	14.3	43.9
Flat-CWA	80.2	39.0	23.9	19.8	44.9
DRAP	**80.3**	42.6	**24.9**	**20.2**	**46.2**

As shown in Table VI, we find that our method achieves higher transferability than Flat-RAP and Flat-CWA. It supports our hypothesis that for a set of surrogate models with distinct loss landscapes, a globally calculated $\boldsymbol{\epsilon}$ could not help to find a flat local minimum of surrogate adversarial risk, as it fails to orient $\hat{\boldsymbol{x}}$ to the real worst-case local neighborhood of each model. Consequently, the crafted $\hat{\boldsymbol{x}}$ will locate at sharp regions of the target model's landscape, slight changes in the loss landscape will cause a significant increase in attack loss.

2) On the Model Discrepancy Penalty

Can within-distribution diversity boost transferability, and if so, how efficiently? To evaluate the impact of incorporating within-distribution diversity when optimizing flatness on narrowing the surrogate-target model discrepancy and controlling the transferability gap, we conduct an ablation study. Specifically, we vary the parameter n, which directly controls the number of diverse surrogate models sampled from each surrogate model components during generating AEs, and thus determines the level of within-distribution diversity in the surrogate model set. We range n from 0 to 40 with a granularity of 5, keeping all other hyper-parameters consistent with Section VI-A. When $n = 5$, n_{LS} is set to 0. Larger values of n correspond to stronger within-distribution diversity, whereas smaller values gradually diminish this influence. At $n = 0$, within-distribution diversity vanishes entirely, and our method reduces to locating an AE that resides in the flat regions of the loss landscapes of the five pretrained surrogate models. As shown in Figure 1, on the whole target model sets, increasing within-distribution diversity consistently improves attack performance over the baseline case ($n = 0$), with peak performance observed at $n = 40$. This result convincingly validates the idea that encouraging AEs to locate within flat regions of the loss landscapes across diverse models increases the likelihood of their generalization to flat regions in unseen models. Consequently, slight changes in the loss landscape are less likely to cause a significant increase in attack loss, thereby improving the transferability of the attack.

However, to find a reverse perturbation, DRAP takes $1 + T$ forward and backward calculations in each iteration. As a higher within-distribution diversity requires more iterations, one may wonder the computational efficiency of the proposed method. The number of gradient calculations N_g as a function of iteration numbers n_{iter} of different methods is summarized in Table XI in *Appendix*. To conduct a fair comparison on computational efficiency with other methods, we evaluate our method as well as others under iterations from 25 to 200 and report the attack performance in Figure 1. We note that

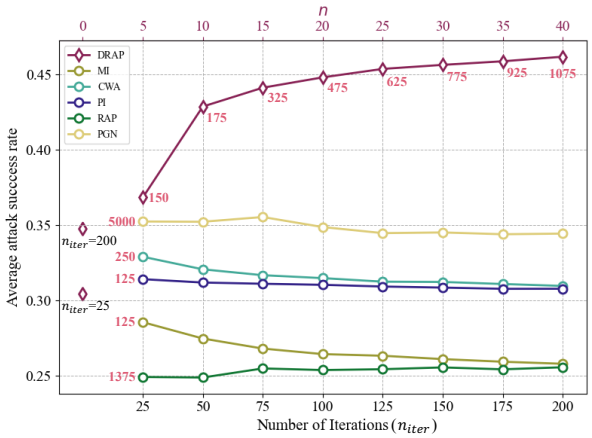

Fig. 1. The average attack success rate of different methods with respect to different number of iterations. The two isolated points indicate our attack with no within-distribution diversity, $i.e.$, $n = 0$, with differnet iteration numbers. The pink numbers represent the number of gradient calculation required for each method under specific iteration numbers.

an iteration of 25 on DRAP is enough to provide significant performance gain over others, while the computational cost is slightly higher than MI-FGSM and PI-FGSM but much lower than others. Meanwhile, with sufficient computational resources to conduct more iterations, DRAP still outperforms compared methods. The results indicate that the proposed DRAP enables a practical trade-off between efficiency and attack performance.

One can observe from Figure 1 that the attack performance of MI-FGSM, PI-FGSM and CWA—methods that accumulate gradients at each iteration—will not benefit from more iterations. Careful tuning of the iteration number is crucial for them, as excessive iterations can lead to gradient overaccumulation, which negatively impacts transferability [27], [36], [42]. In contrast, momentum better synergizes with our method, stabilizing the update directions that vary significantly across diverse surrogate models. Therefore, more iterations ($i.e.$, more diverse surrogate models), better transferability.

Can between-distribution diversity boost transferability, and if so, how efficiently? We have shown that the high within-distribution diversity could effectively and efficiently improve transferability. Here we further validate the necessity of another dimension of model diversity, between-distribution diversity. Specifically, we generate AEs by our method while excluding each prototypical model sets and test the attack performance on models belonging to the ablated prototype, thereby weakening the between-distribution diversity. As shown in Table VII, the average attack success rates on the unseen prototypical sets are 47.8%, 35.2%, 23.7% and 16.6%. Compared to the results before ablation (denoted "Oracle"), we can observe that when the surrogate distributional components are diverse enough to approximate the adversarial vulnerabilities of target models, the attack success rates could be boosted by 32.5%, 7.4%, 1.2% and 3.6%, supporting that the between-distribution diversity plays an important role in generating more transferable AEs. Moreover, we observe that the specific surrogate model architecture chosen to represent each prototype is less significant in our method. This suggests that the between-distribution diversity among prototypes—a higher-level concept than architectural diversity—is more efficient in improving transferability (see *Appendix* D-B for details).

TABLE VII
ATTACK SUCCESS RATES (%, ↑) ON VARIOUS SHIFTED TARGET MODEL SETS. "SHIFTED" MEANS ON EACH RUN WITH A TARGET SET, ITS BELONGING PROTOTYPE WILL BE REMOVED FROM SURROGATE MODELS TO EVALUATE TRANSFERABILITY UNDER SURROGATE-TARGET SHIFT. OTHER EXPERIMENTAL PROTOCOL IS THE SAME AS IN SECTION VI-A. "ORACLE" DENOTES WITH FULL PROTOTYPES IN SURROGATE MODELS.

Attack	Shifted target sets				Avg shifts
	ConvNet Set	Metaformer Set	ConvNet (AT) Set	Metaformer (AT) Set	
MI	32.9	20.7	15.8	12.2	20.4
PI	33.4	21.3	15.8	12.3	20.7
RAP	28.3	16.6	14.7	11.7	17.8
CWA	37.1	23.5	20.5	14.6	23.9
PGN	32.1	23.5	15.9	12.1	20.9
DRAP	**47.8**	**35.2**	**23.7**	**16.6**	**30.8**
DRAP(Oracle)	80.2(+32.5)	42.6(+7.4)	24.9(+1.2)	20.2(+3.6)	-

However, given the endless evolution of model architectures, achieving the ideal between-distribution diversity, as in Oracle, remains challenging. This raises concerns about whether DRAP could maintain its superiority when AEs are expected to transfer to unseen prototypes with unexpected adversarial vulnerabilities in practice. To this end, we build a more strict attack scenario to test transferability under unknown surrogate-target shift in their belonging prototypes, rather than focusing solely on transferability within "training" prototypes explored in Section VI-A. With four prototypes, we choose target models from one prototype and craft AEs on surrogate models from the remaining three prototypes. For instance, when the shifted target models are normally trained convnets, we craft AEs on ViT, XCiT-S(AT) and ResNet-50(AT). We compare the transferability of AEs crafted by DRAP with other five baseline attacks and summarize the results in Table VII. Our methods consistently outperforms others across various surrogate-target shifts, indicating that choosing diverse surrogate componential distributions, albeit imperfect, is still efficiently enough to simulate possible surrogate-target shift in loss landscape. Consequently, seeking flat minima over these distributions significantly narrows the transferability gap.

To analyze the effect of each dimension of model diversity ($i.e.$, within-distribution and between distribution), we ask the above questions. From the results, we find that both dimensions are useful for enhancing the transferability of AEs, as they both narrow the surrogate-target model discrepancy term ℓ_{dis} from different levels. Hence, by leveraging them simultaneously, our method significantly decrease the transferability gap.

VII. CONCLUSION

In this paper, we first prove a bound that provides a guarantee on transferability error. We show that our bound builds a framework generalizing previous approaches and presenting a fresh avenue for the principled analysis of transfer-based attacks. Within our transferability bound, we justify the relationship between flatness and AE transferability and point out the adversarial model discrepancy as another key component to bound transferability. We gain theoretical insights from the derived bound and make algorithmic extensions to our prior

work RAP. The proposed DRAP generates reverse adversarial perturbations tailored to each of the diverse surrogate models, which are selected based on two dimensions of diversity. We conduct extensive experiments on two datasets, covering untargeted and targeted attacks against standard and defense models. We also conduct ablative studies to explore the effect of different penalty terms to verify our theoretical findings.

REFERENCES

[1] F. Yin, Y. Zhang, B. Wu, Y. Feng, J. Zhang, Y. Fan, and Y. Yang, "Generalizable black-box adversarial attack with meta learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Y. Shi, Y. Han, Q. Hu, Y. Yang, and Q. Tian, "Query-efficient black-box adversarial attack with customized iteration and sampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[3] Y. Dong, S. Cheng, T. Pang, H. Su, and J. Zhu, "Query-efficient black-box adversarial attacks guided by a transfer-based prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[4] J. Chen, H. Chen, K. Chen, Y. Zhang, Z. Zou, and Z. Shi, "Diffusion models for imperceptible and transferable adversarial attack," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[5] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *International Conference on Learning Representations*, 2019.

[6] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[7] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Conference on Computer Vision and Pattern Recognition*, 2018.

[8] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Conference on Computer Vision and Pattern Recognition*, 2021.

[9] H. Chen, Y. Zhang, Y. Dong, X. Yang, H. Su, and J. Zhu, "Rethinking model ensemble in transfer-based adversarial attacks," *arXiv e-prints*, 2023.

[10] J. Birrell, M. A. Katsoulakis, and Y. Pantazis, "Optimizing variational representations of divergences and accelerating their statistical estimation," *IEEE Transactions on Information Theory*, 2022.

[11] Z. Qin, Y. Fan, Y. Liu, L. Shen, Y. Zhang, J. Wang, and B. Wu, "Boosting the transferability of adversarial attacks with reverse adversarial perturbation," in *Advances in Neural Information Processing Systems*, 2022.

[12] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[13] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, 2018.

[14] X. Wang, J. Lin, H. Hu, J. Wang, and K. He, "Boosting adversarial transferability through enhanced momentum," *arXiv e-prints*, 2021.

[15] Z. Ge, H. Liu, W. Xiaosen, F. Shang, and Y. Liu, "Boosting adversarial transferability by achieving flat local maxima," in *Advances in Neural Information Processing Systems*, 2023.

[16] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv e-prints*, 2020.

[17] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in *International Conference on Computer Vision*, 2019.

[18] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *International Conference on Computer Vision*, 2021.

[19] J. Zhang, W. Wu, J.-t. Huang, Y. Huang, W. Wang, Y. Su, and M. R. Lyu, "Improving adversarial transferability via neuron attribution-based attacks," in *Conference on Computer Vision and Pattern Recognition*, 2022.

[20] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *International Conference on Computer Vision*, 2021.

[21] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, "Frequency domain model augmentation for adversarial attack," in *European Conference on Computer Vision*, 2022.

[23] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in *International Conference on Computer Vision*, 2023.

[24] Y. Zhu, Y. Chen, X. Li, K. Chen, Y. He, X. Tian, B. Zheng, Y. Chen, and Q. Huang, "Toward understanding and boosting adversarial transferability from a distribution perspective," *Transactions on Image Processing*, 2022.

[25] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," in *Association for the Advancement of Artificial Intelligence*, 2020.

[26] Q. Li, Y. Guo, W. Zuo, and H. Chen, "Making substitute models more bayesian can enhance transferability of adversarial examples," in *International Conference on Learning Representations*, 2023.

[27] Y. Xiong, J. Lin, M. Zhang, J. E. Hopcroft, and K. He, "Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability," in *Conference on Computer Vision and Pattern Recognition*, 2022.

[28] D. A. McAllester, "Pac-bayesian model averaging," in *Proceedings of the twelfth annual conference on Computational learning theory*, 1999.

[29] N. S. Chatterji, B. Neyshabur, and H. Sedghi, "The intriguing role of module criticality in the generalization of deep networks," *arXiv e-prints*, 2019.

[30] I. Csiszár, "On information-type measure of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, 1967.

[31] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, 2016.

[32] R. Agrawal and T. Horel, "Optimal bounds between f-divergences and integral probability metrics," *Journal of Machine Learning Research*, vol. 22, no. 128, pp. 1–59, 2021.

[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[34] N. S. Keskar, J. Nocedal, P. T. P. Tang, D. Mudigere, and M. Smelyanskiy, "On large-batch training for deep learning: Generalization gap and sharp minima," in *International Conference on Learning Representations*, 2017.

[35] M. Gubri, M. Cordy, M. Papadakis, Y. L. Traon, and K. Sen, "Lgv: Boosting adversarial example transferability from large geometric vicinity," in *European Conference on Computer Vision*, 2022.

[36] M. Zheng, X. Yan, Z. Zhu, H. Chen, and B. Wu, "Blackboxbench: A comprehensive benchmark of black-box adversarial attacks," *arXiv e-prints*, 2023.

[37] K. Mahmood, R. Mahmood, and M. Van Dijk, "On the robustness of vision transformers to adversarial examples," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[38] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Conference on Computer Vision and Pattern Recognition*, 2022.

[39] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of dnns," *Advances in neural information processing systems*, 2018.

[40] M. Stephan, M. D. Hoffman, D. M. Blei *et al.*, "Stochastic gradient descent as approximate bayesian inference," *Journal of Machine Learning Research*, 2017.

[41] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, 2013.

[42] Z. Zhao, H. Zhang, R. Li, R. Sicre, L. Amsaleg, M. Backes, Q. Li, and C. Shen, "Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights," *arXiv e-prints*, 2023.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[44] A. Krizhevsky, "Learning multiple layers of features from tiny images," *University of Toronto*, 2009.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[46] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv e-prints*, 2014.

[47] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition*, 2017.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition*, 2015.
[49] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *European Conference on Computer Vision*, 2018.
[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Conference on Computer Vision and Pattern Recognition*, 2018.
[51] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *International Conference on Computer Vision*, 2019.
[52] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Conference on Computer Vision and Pattern Recognition*, 2019.
[53] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019.
[54] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Conference on Computer Vision and Pattern Recognition*, 2022.
[55] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv e-prints*, 2020.
[56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021.
[57] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *International Conference on Computer Vision*, 2021.
[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *International Conference on Computer Vision*, 2021.
[59] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," in *Advances in Neural Information Processing Systems*, 2021.
[60] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *International Conference on Computer Vision*, 2021.
[61] S. Peng, W. Xu, C. Cornelius, M. Hull, K. Li, R. Duggal, M. Phute, J. Martin, and D. H. Chau, "Robust principles: Architectural design principles for adversarially robust cnns," *arXiv e-prints*, 2023.
[62] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do adversarially robust imagenet models transfer better?" in *Advances in Neural Information Processing Systems*, 2020.
[63] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," *arXiv e-prints*, 2020.
[64] C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, and S. Zheng, "A comprehensive study on robustness of image classification models: Benchmarking and rethinking," *International Journal of Computer Vision*, pp. 1–23, 2024.
[65] N. D. Singh, F. Croce, and M. Hein, "Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models," in *Advances in Neural Information Processing Systems*, 2024.
[66] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv e-prints*, 2017.
[67] E. Debenedetti, V. Sehwag, and P. Mittal, "A light recipe to train robust vision transformers," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on Computer Vision and Pattern Recognition*, 2016.
[69] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than cnns?" in *Advances in Neural Information Processing Systems*, 2021.
[70] Z. Wang and Y. Mao, "On f-divergence principled domain adaptation: An improved framework," *arXiv e-prints*, 2024.
[71] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in applied probability*, 1997.
[72] Y. Zhang, M. Long, J. Wang, and M. I. Jordan, "On localized discrepancy for domain adaptation," *arXiv preprint arXiv:2008.06242*, 2020.
[73] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of statistics*, 2000.
[74] Z. Wang, T. Pang, C. Du, M. Lin, W. Liu, and S. Yan, "Better diffusion models further improve adversarial training," in *International Conference on Machine Learning*, 2023.
[75] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Conference on Computer Vision and Pattern Recognition*, 2017.
[76] S. Zagoruyko, "Wide residual networks," *arXiv e-prints*, 2016.
[77] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016.
[78] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," in *Advances in Neural Information Processing Systems*, 2020.
[79] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems*, 2019.
[80] V. Sehwag, S. Wang, P. Mittal, and S. Jana, "Hydra: Pruning adversarially robust neural networks," in *Advances in Neural Information Processing Systems*, 2020.
[81] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations*, 2019.
[82] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning*, 2020.
[83] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, 2019.
[84] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv e-prints*, 2019.
[85] L. Huang, C. Zhang, and H. Zhang, "Self-adaptive training: beyond empirical risk minimization," in *Advances in Neural Information Processing Systems*, 2020.

VIII. Biography Section

Meixi Zheng received the bachelor's degree and the master's degree from the Xidian University in 2019 and 2022. She is currently working toward the PhD degree with the School of Data Science, The Chinese University of Hong Kong, Shenzhen, supervised by Prof. Baoyuan Wu. Her research interest includes adversarial machine learning.

Kehan Wu received the bachelor's degree from Southwest Jiaotong University in 2023. she is currently pursuing a research-based master's degree in the School of Computer and Information Engineering at The Chinese University of Hong Kong under the supervision of both Prof. Rui Huang and Prof. Baoyuan Wu. Her Mphil research focused on transfer-based black-box adversarial attacks.

Yanbo Fan is a Research Scientist at Ant Group. During 2018 to 2023, he worked as a Senior Research Scientist at Tencent AI Lab, Shenzhen, China. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2018, and his B.S. degree in Computer Science and Technology from Hunan University in 2013. His research interests are generative and trustworthy AI.

Rui Huang (Member, IEEE) received his B.Sc. degree from Peking University in 1999, his M.Eng. from the Chinese Academy of Sciences in 2002, and his Ph.D. from Rutgers University in 2008. After completing a postdoctoral appointment at Rutgers, he joined Huazhong University of Science and Technology as a faculty member in 2010. From 2012 to 2016, he was a researcher at NEC Laboratories China. He is currently an Associate Professor at The Chinese University of Hong Kong, Shenzhen. His past research has covered subspace analysis, deformable models, and probabilistic graphical models, with applications in computer vision, pattern recognition, and medical image analysis. His current research interests focus on video analytics, robotic perception and navigation, and autonomous driving. He has authored over 100 papers and led numerous research grants as the principal investigator.

Baoyuan Wu Dr. Baoyuan Wu is a Tenured Associate Professor of School of Data Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China. His research interests are Trustworthy and generative AI, as well as optimization. He is currently serving as Associate Editor of IEEE TIFS and Neurocomputing, and Area Chair of several top-tier AI conferences. He is IEEE Senior Member.

APPENDIX A
PROOFS

A. Technical Lemmas

Lemma 2. *Let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a convex, lower semi-continuous function satisfying $\phi(1) = 0$, and ϕ^* be the Fenchel conjugate function of ϕ, then $\phi^*(\boldsymbol{x}) \geq \boldsymbol{x}$.*

Proof. By definition, $\phi^*(\boldsymbol{x}) = \sup_{\boldsymbol{t}\in \text{dom}\phi} \{\boldsymbol{t}\boldsymbol{x} - \phi(\boldsymbol{t})\} \geq \boldsymbol{t}\boldsymbol{x} - \phi(\boldsymbol{t})$. When $\phi(1) = 0$, we have $\phi^*(\boldsymbol{x}) \geq 1 \cdot \boldsymbol{x} - \phi(1) = \boldsymbol{x}$. □

Introducing the PAC model into transfer-based adversarial attacks to bound the surrogate risk, we have:

Lemma 3. ***(PAC-Bayes [28], [29])*** *For any model distribution $P_{\mathcal{S}}$, prior distribution $\mathcal{P}$ on $\hat{\mathcal{X}}$, $0 < \delta < 1$, with probability $1 - \delta$ over the choice of surrogate model set $\mathcal{M}_{\mathcal{S}} \sim P_{\mathcal{S}}$ with size $K \in \mathbb{N}$, for any distributions $\mathcal{Q}$ on $\hat{\mathcal{X}}$, the following bound holds:*

$$\mathbb{E}_{\hat{\boldsymbol{x}}\sim\mathcal{Q}}[R_{\mathcal{S}}(\hat{\boldsymbol{x}})] \leq \mathbb{E}_{\hat{\boldsymbol{x}}\sim\mathcal{Q}}[R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})] + \sqrt{\frac{KL(\mathcal{Q}||\mathcal{P}) + \log\frac{K}{\delta}}{2(K-1)}}. \tag{A.1}$$

In this bound, the $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$ is the surrogate risk and $R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})$ is the empirical surrogate risk. This PAC-Bayes bound implies: Assuming adversary have enough surrogate model samples, the expected risk of an AE chosen from a distribution $\mathcal{Q}$ can be guaranteed by minimizing the measured loss of distribution $\mathcal{Q}$ and $\frac{KL(\mathcal{Q}\|\mathcal{P})}{n}$, naturally leading to the following optimization method:

1. Fix a distribution $\mathcal{P}$.
2. Collect enough surrogate model instances from $P_{\mathcal{S}}$.
3. Compute the optimal distribution $\mathcal{Q}$ that minimizes the error bound, the right hand side of Equation A.1.
4. Return the crafted AE given by $\mathcal{Q}$.

B. Proof of Theorem 2

In black-box adversarial attacks, it tempting to establish the error bound for the target model including a discrepancy measure (in Definition 4). Consider a loss function $\ell(y_1, y_2)$, such that $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_0^+$. Then we can define a population risk by $R_P(\hat{\boldsymbol{x}}) := \mathbb{E}_{\boldsymbol{w}\sim P}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$. Given two distributions $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$, the following lemmas shows that the difference of risks over $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ can be bounded by the adversarial model discrepancy between $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$. The proof technique we used here is inspired from Wang *et al.* [70].

Theorem 2. (Transferability gap bound) Define $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) = \inf_\alpha \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y) + \alpha)] - \alpha\} - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$. Given the surrogate model distribution $P_{\mathcal{S}}$ and target model distribution $P_{\mathcal{T}}$, for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant $c_1, c_2 \in [0, +\infty)$ subjected to the constraint $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(c_1) \leq c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$, we have

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \frac{1}{c_1}\mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}) + c_2 r. \tag{A.2}$$

Furthermore, if $P_{\mathcal{S}}$ is a mixture distribution of I distributions, *i.e.*, $P_{\mathcal{S}} = \frac{1}{I}\sum_{i\in[I]} P_{\mathcal{S}_i}$, then

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \frac{1}{c_1 I}\sum_{i\in[I]} \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}_i}) + c_2 r. \tag{A.3}$$

Proof. Firstly, $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t)$, which depends on both $t \in \mathbb{R}$ and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$, is defined as follows:

$$\begin{aligned} K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) = &\inf_\alpha \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y) + \alpha)] - \alpha\} \\ &- \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)], \end{aligned}$$

and we define

$$K_{\mathcal{S}}^{r}(t) = \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t). \tag{A.4}$$

We denote for clarity

$$I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) = \inf_\alpha \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\phi^*(t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y) + \alpha)] - \alpha\},$$

which is also the second term of $\mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}})$. Then $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) = I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$

Therefore, for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and $t \in \mathbb{R}$, we have the following inequality holds by Equation A.4:

$$I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] \leq K_{\mathcal{S}}^{r}(t). \tag{A.5}$$

Plugging in $\mathbb{E}_{\mathcal{T}}[t\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$ to the both sides of Equation A.5 and rearranging terms leads to the following inequality:

$$\begin{aligned} &t\left(\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]\right) - K_{\mathcal{S}}^{r}(t) \\ &\leq t\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f_{\hat{\boldsymbol{x}}}, f_{\hat{\boldsymbol{x}}})] - I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) \end{aligned}$$

Since this inequality holds for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and $t \in \mathbb{R}$, we have

$$\begin{aligned} &\sup_{t\in\mathbb{R}} t\left(\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]\right) - K_{\mathcal{S}}^{r}(t) \\ &\leq \sup_{t\in\mathbb{R}} t\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] - I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) \\ &\leq \sup_{t\in\mathbb{R}, \hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} t\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] - I_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(t) \\ &= \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}). \end{aligned}$$

Notice that $\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) = \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)]$. Hence, we have

$$\begin{aligned} &\sup_{t\in\mathbb{R}} t\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) - K_{\mathcal{S}}^{r}(t) \leq \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}) \\ &t\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) - K_{\mathcal{S}}^{r}(t) \leq \mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}) \quad \forall t \in \mathbb{R} \end{aligned} \tag{A.6}$$

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \frac{\mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}) + K_{\mathcal{S}}^{r}(t)}{t} \quad \forall t \geq 0 \tag{A.7}$$

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \inf_{t\geq 0} \frac{\mathrm{D}_\phi^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}}\|P_{\mathcal{S}}) + K_{\mathcal{S}}^{r}(t)}{t}, \tag{A.8}$$

the derivation from Equation A.6 to Equation A.7 restricts the range of t to $t \geq 0$.

Since we assume that there exists constant $c_1, c_2 \in [0, +\infty)$ such that

$$K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(c_1) \leq c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)], \tag{A.9}$$

we have

$$\begin{aligned} K_{\mathcal{S}}^{r}(c_1) &= \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(c_1) \\ &\leq \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}[\ell(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y)] \\ &\leq c_1 c_2 r, \end{aligned}$$

where the last inequality holds by the definition of $\hat{\mathcal{X}}_r$.

Substituting the above inequality into Equation A.8 and replacing t with c_1, we have the following inequality holds for

any c_1, c_2 subject to the constraint A.9:

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \frac{\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})}{c_1} + c_2 r,$$

which completes the proof of Equation A.2.

Then we extend this inequality to the mixture distribution. For this second part, we consider a surrogate mixture of I surrogate distributions where the mixture weight is β_i, denoted as $P_{\mathcal{S}} = \sum_{i\in[I]} \beta_i P_{\mathcal{S}_i}$. Then we can upper bound $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ as follows:

$$\begin{aligned}
&\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) = \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t\in\mathbb{R}} \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right)\right] \\
&- \inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\phi^*\left(t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} \\
&= \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t\in\mathbb{R}} \sum_{i\in[I]} \beta_i \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right)\right] \\
&- \inf_{\alpha\in\mathbb{R}} \sum_{i\in[I]} \beta_i \left(\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}_i}}\left[\phi^*\left(t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\right) \\
&\leq \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t\in\mathbb{R}} \sum_{i\in[I]} \beta_i \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right)\right] \\
&- \sum_{i\in[I]} \beta_i \inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}_i}}\left[\phi^*\left(t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} \qquad \text{(A.10)} \\
&= \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t\in\mathbb{R}} \sum_{i\in[I]} \beta_i (\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right)\right] \\
&- \inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}_i}}\left[\phi^*\left(t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\}) \\
&\leq \sum_{i\in[I]} \beta_i \sup_{\hat{\boldsymbol{x}}' \in \hat{\mathcal{X}}_r, t\in\mathbb{R}} \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right)\right] \\
&- \inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}_i}}\left[\phi^*\left(t \cdot \ell\left(f\left(\hat{\boldsymbol{x}}', \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} \qquad \text{(A.11)} \\
&= \sum_{i\in[I]} \beta_i \mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}_i}),
\end{aligned}$$

where Equation A.10 is by the superadditivity of $\inf$ function and Equation A.11 is by the subadditivity of $\sup$ function. Due to the unavailability of target models, we simply assume the surrogate distributions are equally contributed and set $\beta_i = \frac{1}{I}$. By applying this upper bound of $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ to Equation A.2, we complete the proof of Equation A.3.

□

C. Proof of Corollary 1

Corollary 1. Suppose $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$. Given the surrogate model distribution $P_{\mathcal{S}}$ and target model distribution $P_{\mathcal{T}}$, for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ and constant c_1 satisfying $0 \leq c_1 \leq 1$, we have

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \frac{1}{c_1} \mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}),$$

where $\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) = \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} | \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] |$.

Proof. Here, we instantiate the bound in Theorem 2 with TV distance. Specifically, let $\phi_{\text{TV}}(u) = | u - 1 |$, its convex conjugate function is $\phi_{\text{TV}}^*(v) = v$, where v takes value in $[-1, 1]$.

$$\begin{aligned}
&\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) \\
&= \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r, -1\leq t\leq 1} \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] - \\
&\inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\phi_{\text{TV}}^*\left(t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} \\
&\geq t\left(\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right]\right) \qquad \text{(A.12)} \\
&= t\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}). \qquad \text{(A.13)}
\end{aligned}$$

The above inequality A.12 holds for any $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r, -1 \leq t \leq 1$. Since when $t = 0$, this holds by the non-negativity of $\mathrm{D}_{\phi}^{\hat{\mathcal{X}}_r}$ discrepancy, and when $0 < t \leq 1$, we have

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \inf_{0<t\leq 1} \frac{\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})}{t}.$$

Overall, we have

$$\mathcal{E}_{\text{trans}}(\hat{\boldsymbol{x}}) \leq \inf_{0\leq t\leq 1} \frac{\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})}{t}.$$

Substituting t with c_1 gives the desired results.

We further simplify $\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}})$ as follows

$$\begin{aligned}
&\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}(P_{\mathcal{T}} \| P_{\mathcal{S}}) \\
&= \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r, -1\leq t\leq 1} \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] - \\
&\inf_{\alpha\in\mathbb{R}} \{\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\phi_{\text{TV}}^*\left(t\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right) + \alpha\right)\right] - \alpha\} \\
&= \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r, -1\leq t\leq 1} t(\mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] \\
&- \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right]) \qquad \text{(A.14)} \\
&= \sup_{\hat{\boldsymbol{x}}\in\hat{\mathcal{X}}_r} | \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{T}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] - \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f\left(\hat{\boldsymbol{x}}, \boldsymbol{w}\right), y\right)\right] |.
\end{aligned}$$

The Equation A.14 is by the fact that $\sup_{t\in[-1,1]} t \cdot a = |a|$. The above derivation recovers the integral probability metric form of TV distance defined by [71]. □

The above lemma instantiates the bound in Theorem 2 and gives a clearer clue for the condition $c_1, c_2 \in [0, +\infty)$ subjected to the constraint $K_{\mathcal{S}}^{\hat{\boldsymbol{x}}}(c_1) \leq c_1 c_2 \mathbb{E}_{\boldsymbol{w}\sim P_{\mathcal{S}}}\left[\ell\left(f(\hat{\boldsymbol{x}}, \boldsymbol{w}), y\right)\right]$. In the case of TV, we explicitly show that the constraint simplifies to $0 \leq c_1 \leq 1, c_2 = 0$. This bound shares a similar form as the domain adaptation bounds in Theorem 2 in [72] when setting $c_1 = 1$, without violating this bound. While our task does not require a separate ideal joint error term; instead, this term is implicitly included within the discrepancy term. Since the $\mathrm{D}_{\text{TV}}^{\hat{\mathcal{X}}_r}$-based transferability gap bound is merely a special case of our main result in Theorem 2, our result provides a more general framework that encompasses the family of ϕ-divergences.

D. Proof of Theorem 3

Below, we show an empirical estimation of the surrogate adversarial risk $R_{\mathcal{S}}(\hat{\boldsymbol{x}})$. The proof technique we used here is inspired from Foret *et al.* [16] and Chatterji *et al.* [29].

Theorem 3. (Surrogate risk bound) For any $\rho > 0, 0 < \delta < 1$, model distribution $P_{\mathcal{S}}$, and $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$, with probability $1 - \delta$ over the choice of surrogate model set $\mathcal{M}_{\mathcal{S}} \sim P_{\mathcal{S}}$ with size

$K \in \mathbb{N}$, we have

$$R_{\mathcal{S}}(\hat{\boldsymbol{x}}) \leq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon}) + \sqrt{\frac{\frac{d}{2}\log(1 + \frac{\gamma^2}{\rho^2}(1 + \sqrt{\frac{\log K}{d}})^2) + \log\frac{K}{\delta} + \tilde{\mathcal{O}}(1)}{2(K-1)}}. \tag{A.15}$$

where $\tilde{\mathcal{O}}(1)$ term equals to $\varepsilon = \frac{1}{2} + 2\log(2 + 3d + 6r^2K + 4d\log(\sqrt{d} + \sqrt{logK}))$.

Proof. For $\hat{\boldsymbol{x}} \in \hat{\mathcal{X}}_r$ with $\hat{\mathcal{X}}_r = \left\{\hat{\boldsymbol{x}} \in \hat{\mathcal{X}} \mid R_{\mathcal{S}}(\hat{\boldsymbol{x}}) \leq r\right\}$, if r is a relatively small value, then we assume without loss of generality that adding a noise $\boldsymbol{\epsilon}$ following the Gaussian distribution, *i.e.*, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, on $\hat{\boldsymbol{x}}$ will not further reduce its surrogate adversarial risk. That is to say,

$$R_{\mathcal{S}}(\hat{\boldsymbol{x}}) \leq \mathbb{E}_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[R_{\mathcal{S}}(\hat{\boldsymbol{x}} + \boldsymbol{\epsilon})]. \tag{A.16}$$

From Lemma 3, we have that for any model distribution $P_{\mathcal{S}}$, $n \in \mathbb{N}$, prior distribution $\mathcal{P}$ over AEs, $0 < \delta < 1$, with probability $1 - \delta$ over the choice of set $\mathcal{M}_{\mathcal{S}} \sim P_{\mathcal{S}}$ with size K, for any posterior distributions $\mathcal{Q}$ over AEs, the following bound holds:

$$\mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathcal{Q}}[R_{\mathcal{S}}(\hat{\boldsymbol{x}})] \leq \mathbb{E}_{\hat{\boldsymbol{x}} \sim \mathcal{Q}}[R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}})] + \sqrt{\frac{\mathrm{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{K}{\delta}}{2(K-1)}}. \tag{A.17}$$

We first consider the KL divergence term in Equation A.17. Let the prior $\mathcal{P}$ be a Gaussian distribution centered at the benign image $\boldsymbol{x}$ with covariance matrix $\sigma_p^2\boldsymbol{I}$, *i.e.*, $\mathcal{P} = \mathcal{N}(\boldsymbol{x}, \sigma_p^2\boldsymbol{I})$. Let posterior $\mathcal{Q}$ be a Gaussian distribution centered at $\hat{\boldsymbol{x}}$ which is the benign image $\boldsymbol{x}$ with an additive adversarial perturbation $\boldsymbol{\xi}$ — $\boldsymbol{x}+\boldsymbol{\xi}$, with covariance matrix $\sigma^2\boldsymbol{I}$, *i.e.*, $\mathcal{Q} = \mathcal{N}(\boldsymbol{x}+\boldsymbol{\xi}, \sigma^2\boldsymbol{I})$, as if we have added the noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Then KL divergence term can be written as:

$$\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) = \frac{1}{2}\left[\frac{d\sigma^2 + \|\boldsymbol{\xi}\|_2^2}{\sigma_p^2} - d + d\log\left(\frac{\sigma_p^2}{\sigma^2}\right)\right]. \tag{A.18}$$

The variance of prior σ_p is selected to minimize the above KL divergence. Given σ, we differentiate $\mathrm{KL}(\mathcal{Q}\|\mathcal{P})$ with respect to σ_p and set the derivative to zero, solving

$$\sigma_p^{*2} = \sigma^2 + \|\boldsymbol{\xi}\|_2^2/d.$$

However, the prior is selected in advance and independent of the training data, which the posterior depends on. Hence, the above solution is not allowed. However, one can optimize the prior standard deviation σ_p over a pre-defined set and use a union bound augment to get the bound for best σ_p in this set. Let the pre-defined set be $\{c\exp((1-j)/d) | j \in \mathbb{N}\}$. If for each j one chooses σ_p to be

$$\sigma_p = c\exp((1-j)/d),$$

such that the above PAC-Bayes bound holds with probability $1 - \delta_j$ where $\delta_j = \frac{6\delta}{\pi^2 j^2}$, then all bound can be combined according to union bound and hold with probability

$$1 - \sum_{j=1}^{\infty}\delta_j = 1 - \sum_{j=1}^{\infty}\frac{6\delta}{\pi^2 j^2} = 1 - \delta.$$

As the right hand side of Equation A.15 is lower bounded by $\sqrt{\frac{d\log\left(1+\frac{\|\boldsymbol{\xi}\|_2^2}{\rho^2}\right)}{4K}}$ since $\boldsymbol{\xi}$ is restricted in the γ-norm ball. When $\sqrt{\frac{d\log\left(1+\frac{\|\boldsymbol{\xi}\|_2^2}{\rho^2}\right)}{4K}} > r$, this bound holds trivially. Thus we only consider when $\sqrt{\frac{d\log\left(1+\frac{\|\boldsymbol{\xi}\|_2^2}{\rho^2}\right)}{4K}} \leq r$, leading to

$$\|\boldsymbol{\xi}\|_2^2 \leq \rho^2\left(\exp\left(\frac{4r^2K}{d}\right) - 1\right) \tag{A.19}$$

Therefore, by Equation A.19, we have

$$\begin{aligned}\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d &\leq \sigma^2 + \frac{\rho^2}{d}\left(\exp\left(\frac{4r^2K}{d}\right) - 1\right) \\ &\leq \sigma^2 + \rho^2\exp\left(\frac{4r^2K}{d}\right).\end{aligned} \tag{A.20}$$

As $\sigma_p = c\exp((1-j)/d)$, we consider the case $j = 1 - d\log(\sigma_p^2/c) = \left\lfloor 1 - d\log\left(\left(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d\right)/c\right)\right\rfloor$, where c can be set as $\sigma^2 + \rho^2\exp\left(\frac{4r^2K}{d}\right)$to make sure $j \in \mathbb{N}$ according to Equation A.20. For this j, we have

$$\begin{aligned}1 - d\log(\sigma_p^2/c) &\leq 1 - d\log\left(\left(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d\right)/c\right) \\ \sigma_p^2 &\geq \sigma^2 + \|\boldsymbol{\xi}\|_2^2/d.\end{aligned}$$

Similarly, we can derive an upper bound for σ_p^2,

$$\begin{aligned}1 - d\log(\sigma_p^2/c) &\geq 1 - d\log\left(\left(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d\right)/c\right) - 1 \\ \sigma_p^2 &\leq \exp(\frac{1}{d})(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d).\end{aligned}$$

With the above lower bound and upper bound, the KL term in Equation A.18 can be written as:

$$\begin{aligned}&\mathrm{KL}(\mathcal{Q}\|\mathcal{P}) = \frac{1}{2}\left[\frac{d\sigma^2 + \|\boldsymbol{\xi}\|_2^2}{\sigma_p^2} - d + d\log\left(\frac{\sigma_p^2}{\sigma^2}\right)\right] \\ &\leq \frac{1}{2}\left[\frac{d\sigma^2 + \|\boldsymbol{\xi}\|_2^2}{\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d} - d + d\log\left(\frac{\exp(\frac{1}{d})(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d)}{\sigma^2}\right)\right] \\ &= \frac{1}{2}\left[1 + d\log\left(\frac{d\sigma^2 + \|\boldsymbol{\xi}\|_2^2}{d\sigma^2}\right)\right].\end{aligned} \tag{A.21}$$

We then consider the log term in Equation A.17. Under the case of above discussed j, the bound holds with probability $1 - \delta_j$ where $\delta_j = \frac{6\delta}{\pi^2 j^2}$. Therefore, the rest part could be simplified as

$$\begin{aligned}&\log\frac{K}{\delta_j} = \log\frac{K}{\delta} + \log\frac{\pi^2 j^2}{6} \\ &\leq \log\frac{K}{\delta} + \log\frac{\pi^2\left(1 + d\log\left(c/\left(\sigma^2 + \|\boldsymbol{\xi}\|_2^2/d\right)\right)\right)^2}{6} \quad \text{(A.22)} \\ &\leq \log\frac{K}{\delta} + \log\frac{\pi^2\left(1 + d\log\left(c/\sigma^2\right)\right)^2}{6} \\ &= \log\frac{K}{\delta} + \log\frac{\pi^2\left(1 + d\log\left(1 + \exp\left(\log(\frac{\rho^2}{\sigma^2}) + \frac{4r^2K}{d}\right)\right)\right)^2}{6} \quad \text{(A.23)} \\ &\leq \log\frac{K}{\delta} + \log\frac{\pi^2\left(1 + 2d + 4r^2K + d\log(\frac{\rho^2}{\sigma^2})\right)^2}{6} \quad \text{(A.24)} \\ &\leq \log\frac{K}{\delta} + 2\log(2 + 3d + 6r^2K + 2d\log(\frac{\rho^2}{\sigma^2})), \quad \text{(A.25)}\end{aligned}$$

where Equation A.22 is by the upper bound of j. Equation A.23 is by the value set for c. A.24 is by the fact that $\log(1 + \mathrm{e}^x) \leq 2 + x$, for every $x \geq 0$.

Substituting the upper bound for KL term in Equation A.21 and the upper bound for log term in Equation A.25, we have

the following PAC-Bayes bound

$$\mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma^2)}[R_{\mathcal{S}}(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon})] \leq \mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma^2)}[R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon})]+ \sqrt{\frac{\frac{d}{2}\log\left(\frac{d\sigma^2+\|\boldsymbol{\xi}\|_2^2}{d\sigma^2}\right)+2\log(2+3d+6r^2K+2d\log(\frac{\rho^2}{\sigma^2}))+\varepsilon'}{2(K-1)}}, \tag{A.26}$$

where $\varepsilon' = \log\frac{K}{\delta}+\frac{1}{2}$.

As we set $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, we have

$$\|\boldsymbol{\epsilon}\|_2^2 = \sum_{i=1}^{d}\epsilon_i^2 = \sigma^2\chi_d^2,$$

where χ_d^2 is a chi-square random variable with d degrees of freedom (mean d, variance $2d$). By Lemma 1 in [73], we have

$$\Pr\left[\chi_d^2 - d \geq 2\sqrt{dt}+2t\right] \leq e^{-t}, \quad t>0.$$

Multiplying both sides by σ^2 yields

$$\Pr\left[\|\boldsymbol{\epsilon}\|_2^2 - d\sigma^2 \geq 2\sigma^2\sqrt{dt}+2\sigma^2 t\right] \leq e^{-t}. \tag{A.27}$$

Set $e^{-t} = 1/\sqrt{K}$, then $t = \log\sqrt{K}$. Substituting t into Equation A.27, we have

$$\Pr\left[\|\boldsymbol{\epsilon}\|_2^2 \leq \sigma^2(d+2\sqrt{d\log\sqrt{K}}+2\log\sqrt{K})\right] \geq 1-1/\sqrt{K}. \tag{A.28}$$

As $d+2\sqrt{d\log\sqrt{K}}+2\log\sqrt{K} \leq d\left(1+\sqrt{\frac{\log K}{d}}\right)^2$, we have, with probability $\geq 1-1/\sqrt{K}$,

$$\|\boldsymbol{\epsilon}\|_2^2 \leq d\sigma^2\left(1+\sqrt{\frac{\log K}{d}}\right)^2. \tag{A.29}$$

By defining $\rho^2 = d\sigma^2\left(1+\sqrt{\frac{\log K}{d}}\right)^2$, Equation A.29 becomes

$$\|\boldsymbol{\epsilon}\|_2^2 \leq \rho^2, \quad \text{with probability } 1-1/\sqrt{K}.$$

Combining Equation A.26 and A.16 and splitting the expectation over the two events $\{\|\boldsymbol{\epsilon}\|_2 \leq \rho\}$ with probability $1-1/\sqrt{K}$ and $\{\|\boldsymbol{\epsilon}\|_2 > \rho\}$ with probability $1/\sqrt{K}$,

$$\begin{aligned}
&R_{\mathcal{S}}(\hat{\boldsymbol{x}}) \leq \mathbb{E}_{\epsilon_i\sim\mathcal{N}(0,\sigma^2)}[R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon})]+ \\
&\sqrt{\frac{\frac{d}{2}\log\left(\frac{d\sigma^2+\|\boldsymbol{\xi}\|_2^2}{d\sigma^2}\right)+2\log(2+3d+6r^2K+2d\log(\frac{\rho^2}{\sigma^2}))+\varepsilon'}{2(K-1)}} \\
&\leq (1-1/\sqrt{K})\max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon}) + 1/\sqrt{K}+ \\
&\sqrt{\frac{\frac{d}{2}\log(1+\frac{\|\boldsymbol{\xi}\|_2^2}{\rho^2}(1+\sqrt{\frac{\log K}{d}})^2)+\varepsilon''+\varepsilon'}{2(K-1)}} \\
&\leq \max_{\|\boldsymbol{\epsilon}\|_2\leq\rho} R_{\hat{\mathcal{S}}}(\hat{\boldsymbol{x}}+\boldsymbol{\epsilon})+ \\
&\sqrt{\frac{\frac{d}{2}\log(1+\frac{\gamma^2}{\rho^2}(1+\sqrt{\frac{\log K}{d}})^2)+\varepsilon''+\varepsilon'}{2(K-1)}},
\end{aligned}$$

where $\varepsilon'' = 2\log(2+3d+6r^2K+4d\log(\sqrt{d}+\sqrt{logK}))$. This completes the proof. □

APPENDIX B
EXPERIMENTS ON CIFAR-10

A. Main Results

Baselines For the experiments conducted on CIFAR-10, we adopt the same set of baseline methods used in the ImageNet experiments.

Models For surrogate models on CIFAR-10, we adopt three architectures ($I=3$) from three prototypical model categories: ResNet-50 [68] from normally trained convnets, ViT [55] from normally trained metaformers, and WideResNet-70-16 (AT) [74] from adversarially trained convnets. Due to the limited availability of adversarially trained metaformer models on CIFAR-10, we exclude them from the surrogate models. For DRAP, model samples are gathered as proposals at each epoch during the fine-tuning of the three pretrained models, which are optimized using their respective training strategies over $n=40$ additional epochs. In order to get more diverse samples, we fine-tune the three pretrained models with larger constant learning rates of 0.05, 0.03, and 0.02, respectively, for ResNet-50, ViT, and WideResNet-70-16 (AT), without significantly degrading their clean accuracy. All compared methods also generate AEs using the logits of the three surrogate models, leveraging the fusion strategy outlined in [7]. To assess the transferability of these AEs, we evaluate them across a diverse set of target models. These target models are grouped into three categories: ConvNet Set, ConvNet(AT) Set, and Metaformer Set, as shown in Tab. VIII.

Implementation Details For both untargeted and targeted attack scenarios on CIFAR-10, the adversarial perturbation magnitude is constrained to $\gamma = 8/255$, with a step size of $\beta_{\hat{\boldsymbol{x}}} = 8/255/10$ for all methods. The RAP method retains its iteration number of 400 to ensure comparable evaluation. For all other methods, a standard iteration count of 120 is used to maintain consistency across experiments. For DRAP, the key hyper-parameters are configured as follows: the number of samples per model distribution $n=40$, the inner optimization iteration count $T=5$, the late start iteration number $n_{LS}=5$, the inner step size $\beta_\epsilon = 0.1/255$, and the decay factor $\mu = 1$. The total number of iterations for generating adversarial examples in DRAP is $n\times I = 120$, aligning with the iteration count of other methods to ensure fairness in comparison. For all baseline methods, we adhere to the protocols established in BlackboxBench to maintain consistency across experimental setups.

Results of Untargeted Attacks Table VIII presents the untargeted attack results on CIFAR-10 across normally trained and adversarially trained target model sets. Overall, our proposed method achieves the highest average attack success rate, surpassing both input-transformation based and other optimization-based methods. This outcome aligns with the ImageNet findings, indicating that our strategy generalizes effectively to a dataset with different characteristics. Taking a closer look, we observe that SIA remains highly competitive against normally trained target models but underperforms significantly on adversarially trained models compared to DRAP. In contrast, our approach maintains strong performance on adversarially trained models without sacrificing success rates on normally trained ones, striking a well-rounded balance across all target sets. This reinforces the idea that simultaneously pursuing flatness among diverse surrogate models can secure robust adversarial transferability, regardless of the target's defense mechanisms. Similar to ImageNet results, CWA continues to exhibit promising effectiveness against adversarially trained targets. However, its overall success rate remains lower, likely

TABLE VIII
UNTARGETED ATTACK SUCCESS RATES (%,↑) ON CIFAR-10 DATASET. THE AES ARE CRAFTED FROM THREE SURROGATE MODELS (RESNET-50, VIT, AND WIDERESNET-70-16(AT)), AGAINST 20 TARGET MODELS FALLING INTO THREE PROTOTYPES (NORMALLY AND ADVERSARIALLY TRAINED CONVNETS AND NORMALLY TRAINED METAFORMERS). **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Target Model Set		I-FGSM	DI2-FGSM	SI-FGSM	Admix	TI-FGSM	SSA	SIA	MI-FGSM	PI-FGSM	VT-FGSM	PGN	CWA	SVRE	RAP	**DRAP**
ConvNet Set	AlexNet [45]	28.4	31.0	29.0	29.5	28.6	33.4	32.5	29.4	30.2	30.3	34.0	35.2	28.2	30.9	**41.5**
	DenseNet [47]	88.1	94.8	91.7	95.2	88.2	96.3	**97.9**	90.4	92.3	95.8	96.5	85.7	88.5	96.5	96.7
	ResNeXt [75]	93.3	97.3	95.8	97.8	93.3	97.7	**99.0**	94.4	94.8	97.3	98.9	88.7	92.3	97.7	97.9
	WRN-28-10-drop [76]	87.6	94.4	91.3	95.1	87.9	96.7	**97.7**	90.7	92.2	96.0	96.8	85.0	87.7	96.5	96.9
	GoogleNet [48]	89.7	95.3	92.8	95.7	89.9	96.8	**98.5**	92.2	93.5	96.3	96.4	86.8	89.3	97.0	96.9
	MobileNetv2 [50]	89.2	96.0	92.4	96.0	89.1	97.2	**98.2**	92.3	93.1	96.2	97.3	86.0	89.1	96.7	97.7
	PreResNet [77]	96.7	98.7	98.0	99.0	96.5	99.0	99.1	97.4	96.8	98.3	99.5	91.8	95.8	98.5	99.2
	Average	81.9	86.8	84.4	86.9	81.9	88.2	89.0	83.8	84.7	87.2	88.5	79.9	81.6	87.7	**89.5**
MetaFormer Set	ViT-T [55]	73.1	87.6	77.6	85.8	72.9	84.0	**93.9**	76.3	79.9	88.1	88.7	75.9	72.4	84.8	80.4
	Swin-S [58]	76.1	90.0	89.6	89.5	89.2	90.1	**95.9**	89.4	82.1	89.9	92.6	77.3	76.3	90.1	83.7
	Swin-B [58]	71.4	89.9	89.6	89.4	88.9	87.0	**93.4**	88.9	78.9	89.6	90.3	74.9	73.2	89.9	79.7
	DeiT-T [56]	74.9	88.8	79.6	87.9	75.2	87.5	**95.0**	78.6	80.9	89.4	90.5	77.5	74.2	85.8	81.0
	DeiT-B [56]	72.0	87.7	76.5	85.1	72.1	83.1	**92.8**	74.1	79.2	88.5	88.4	74.0	72.7	83.9	78.0
	Average	73.5	88.8	82.6	87.5	79.7	86.3	**94.2**	81.4	80.2	89.1	90.1	75.9	73.8	86.9	80.6
ConvNet(AT) Set	PreActResNet-18 [78]	23.6	24.4	23.6	24.1	23.7	26.1	24.7	24.3	23.9	24.4	25.9	29.2	23.3	24.7	**34.0**
	WideResNet-28-10 [79]	14.4	15.0	14.0	14.7	14.4	15.9	14.4	15.1	14.3	15.0	16.0	21.4	14.0	15.0	**26.2**
	WideResNet-28-10 [80]	15.2	16.0	14.9	15.5	15.2	16.7	15.0	15.8	15.2	15.8	16.8	22.0	14.6	15.8	**27.0**
	WideResNet-28-10 [81]	16.4	17.0	16.2	16.6	16.4	18.2	16.7	17.1	16.7	16.9	18.4	22.6	16.0	17.3	**27.4**
	WideResNet-34-20 [82]	18.3	18.9	18.4	18.8	18.5	20.2	18.9	19.0	18.4	18.8	20.3	24.7	18.0	18.9	**29.6**
	WideResNet-34-10 [83]	18.8	19.6	18.8	19.2	18.9	20.6	19.4	19.6	19.1	19.4	20.7	25.5	18.5	19.6	**30.7**
	ResNet-50 [84]	16.9	17.7	16.6	17.4	17.0	18.7	17.3	17.6	17.2	17.5	18.8	23.7	16.7	17.8	**29.1**
	WideResNet-34-10 [85]	19.9	20.6	19.8	20.3	20.1	21.2	20.1	20.5	20.0	20.4	21.3	26.1	19.6	20.4	**30.8**
	Average	17.9	18.6	17.8	18.3	18.0	19.7	18.3	18.6	18.1	18.5	19.8	24.4	17.6	18.7	**29.4**
Overall Average		54.2	60.0	57.3	59.6	55.8	60.3	62.0	57.2	56.9	60.2	61.4	56.7	54.0	59.9	**63.2**

due to its finding a suboptimal universal perturbation.

Results of Targeted Attacks Table IX presents the targeted attack results on CIFAR-10. As with ImageNet, targeted attacks remain particularly challenging, especially on adversarially trained models in the ConvNet(AT) Set. Among all evaluated methods, DRAP demonstrates a clear advantage on the ConvNet(AT) Set, achieving significantly higher success rates compared to other methods, showcasing its effectiveness against robust adversarially trained targets. Beyond the ConvNet(AT) Set, DRAP also performs strongly on the ConvNet Set and MetaFormer Set, achieving competitive success rates that match or surpass other leading methods in many cases. This highlights the adaptability of DRAP across diverse target model categories, maintaining a strong balance between attacking normally trained and adversarially trained models. Overall, these results reinforce the robustness and competitiveness of DRAP in targeted attack scenarios, particularly when faced with highly defensive models.

B. Composition with Input-Transformation Based Attacks

Following the combination experiments on ImageNet, we conduct similar tests on CIFAR-10 by integrating DRAP with several well-known input-transformation based attacks and optimization-based attacks. Due to computational constraints, we omit the combination of PGN and SIA. The experimental setup follows the untargeted protocol as described earlier. Table X summarizes the results on CIFAR-10. Consistent with the findings on ImageNet, integrating DRAP with input-transformation based attacks significantly enhances their base performance, leading to superior attack success rates across both normal and adversarially trained models. Notably, our combined method achieves the highest attack performance in every combination. It also outperforms all other methods on the most challenging ConvNet(AT) Set. Furthermore, it demonstrates competitive performance on the ConvNet Set and MetaFormer Set, maintaining a strong balance across different target model categories. These results validate the scalability and adaptability of DRAP on CIFAR-10. Not only does it excel as a standalone approach, but it also amplifies the effectiveness of input-transform methods, confirming its potential for achieving state-of-the-art adversarial transferability.

APPENDIX C
FLATNESS OPTIMIZATION ALGORITHMS

The details of using strategies from RAP and CWA to optimize the loss sharpness over diverse surrogate models are shown in Algorithm 2 and Algorithm 3, respectively. We stick to their original algorithms but substitute the surrogate models seen at each iteration j with a batch of models from the diverse model set. Specifically, we sample one model weight from every model architecture to compose one batch. When fusing the gradients of multiple models, we use the logits ensemble strategy as suggested in RAP and CWA. For hyper-parameters, we set perturbation budget $\gamma = 4/255$ with step size $\alpha = 2/255$, $n = 40$ and follow the optimal settings reported in their papers to set their own hyper-parameters. In Flat-RAP, we set step size $\beta = \alpha$ and $n_{LS} = 5$, which is same as in DRAP. In Flat-CWA, we set decay factor $\mu = 1$, step sizes $\beta = 50$, $r = \gamma/15$.

APPENDIX D
ADDITIONAL ABLATIVE STUDY

A. Parameter Study on n_{LS}

As the late start iteration number n_{LS} decides when the sharpness penalty begins to take effect during the optimization process, we range n_{LS} from 0 to n with $n_{LS} = \{0, 5, 15, 25, 35, 40\}$. When $n_{LS} = 0$, the sharpness penalty is active throughout the entire optimization process. As n_{LS} increases, the influence of the sharpness penalty gradually weakens, eventually vanishing at $n_{LS} = K$, at which point DRAP reduces to a standard model ensemble attack. As shown

TABLE IX
TARGETED ATTACK SUCCESS RATES (%,↑) ON CIFAR-10 DATASET. THE AES ARE CRAFTED FROM THREE SURROGATE MODELS (RESNET-50, VIT, AND WIDERESNET-70-16(AT)), AGAINST 20 TARGET MODELS FALLING INTO THREE PROTOTYPES (NORMALLY AND ADVERSARIALLY TRAINED CONVNETS AND NORMALLY TRAINED METAFORMERS). **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Target Model Set		I-FGSM	DI2-FGSM	SI-FGSM	Admix	TI-FGSM	SSA	SIA	MI-FGSM	PI-FGSM	VT-FGSM	PGN	CWA	SVRE	RAP	**DRAP**
ConvNet Set	AlexNet [45]	3.4	4.3	3.5	3.8	3.3	5.1	4.9	3.8	4.0	3.8	5.4	6.2	3.2	4.2	**8.6**
	DenseNet [47]	53.3	74.3	64.1	79.7	53.0	83.0	**86.0**	56.6	64.3	82.3	78.2	59.8	54.6	72.5	84.8
	ResNeXt [75]	66.1	83.4	75.4	85.7	65.9	84.8	**91.9**	68.4	68.8	87.6	80.6	63.2	63.2	76.4	86.0
	WRN-28-10-drop [76]	55.8	75.5	67.3	81.0	55.3	81.9	**84.6**	60.9	67.3	84.3	79.8	60.5	55.6	74.3	82.1
	GoogleNet [48]	55.5	76.0	65.9	80.8	55.2	85.2	**91.4**	60.3	66.5	84.0	80.6	60.0	55.1	73.5	86.2
	MobileNetv2 [50]	52.4	74.6	63.2	79.1	52.4	81.0	**84.6**	57.0	65.0	81.7	80.0	59.5	53.9	72.1	82.0
	PreResNet [77]	73.2	87.3	83.1	92.5	73.0	89.9	**92.6**	74.1	73.6	92.1	85.4	68.0	72.3	79.4	90.6
	Average	51.4	67.9	60.4	71.8	51.2	73.0	**76.6**	54.4	58.5	73.7	70.0	53.9	51.1	64.6	74.3
MetaFormer Set	ViT-T [55]	33.3	57.1	40.3	56.1	33.1	47.6	**58.2**	36.7	44.9	56.4	42.1	44.4	33.1	51.9	52.9
	Swin-S [58]	36.9	14.7	44.2	63.2	36.9	57.1	**66.9**	12.7	46.5	14.9	51.1	47.3	38.0	14.1	58.3
	Swin-B [58]	33.0	14.3	39.3	58.2	33.0	54.3	**59.6**	12.2	45.1	14.2	46.7	45.7	35.6	13.8	54.8
	DeiT-T [56]	34.6	59.5	41.0	54.5	34.7	51.7	**61.3**	38.7	44.5	59.7	47.1	45.4	33.6	53.2	53.1
	DeiT-B [56]	33.8	62.2	40.7	59.2	34.5	48.6	**60.8**	36.1	45.4	58.9	42.1	43.9	34.4	53.0	52.2
	Average	34.3	41.6	41.1	58.2	34.4	51.9	**61.4**	27.3	45.3	40.8	45.8	45.3	34.9	37.2	54.3
ConvNet(AT) Set	PreActResNet-18 [78]	3.0	3.2	2.9	3.1	3.0	3.6	3.0	3.2	3.1	2.9	3.6	4.8	2.6	3.3	**6.4**
	WideResNet-28-10 [79]	2.2	2.4	2.0	2.2	2.2	2.6	2.1	2.4	2.1	2.0	2.4	4.6	1.7	2.5	**6.1**
	WideResNet-28-10 [80]	2.4	2.5	2.2	2.5	2.4	2.9	2.2	2.5	2.4	2.3	2.8	4.7	1.8	2.7	**6.4**
	WideResNet-28-10 [81]	2.4	2.5	2.2	2.3	2.3	2.7	2.2	2.5	2.3	2.2	2.7	4.5	1.8	2.5	**6.0**
	WideResNet-34-20 [82]	3.0	3.2	2.9	3.1	3.0	3.4	3.0	3.2	2.9	2.9	3.4	5.1	2.5	3.3	**6.5**
	WideResNet-34-10 [83]	2.7	2.9	2.6	2.8	2.7	3.2	2.8	2.9	2.7	2.6	3.1	4.9	2.3	3.0	**6.5**
	ResNet-50 [84]	2.4	2.6	2.2	2.5	2.4	2.9	2.4	2.7	2.4	2.3	2.9	4.9	2.0	2.7	**7.0**
	WideResNet-34-10 [85]	2.7	3.0	2.6	2.8	2.8	3.1	2.7	3.0	2.7	2.7	3.1	4.8	2.4	3.0	**6.3**
	Average	2.6	2.8	2.4	2.6	2.6	3.0	2.5	2.8	2.5	2.4	3.0	4.8	2.1	2.9	**6.4**
Overall Average		27.6	35.3	32.4	40.8	27.6	39.7	**43.2**	27.0	32.8	37.0	37.2	32.1	27.5	33.1	42.1

TABLE X
ATTACK SUCCESS RATES ON CIFAR-10 DATASET (%, ↑) OF MI, PI, VT, RAP, PGN, CWA, AND DRAP WHEN INTEGRATED WITH DI, TI, ADMIX, AND SIA, RESPECTIVELY. THE INDENTATION DENOTES COMBINATION. THE RESULTS ARE AVERAGED ON EACH MODEL SET.

Attack	ConvNet Set	MetaFormer Set	CNN (AT) Set	*Overall Average*
DI-FGSM	86.7	88.8	18.6	60.0
+ MI	88.7	91.7	19.8	61.9
+ PI	89.0	91.9	19.9	62.1
+ VT	89.2	94.3	19.6	62.6
+ RAP	86.3	82.0	19.3	58.4
+ PGN	**89.8**	**92.3**	20.1	62.6
+ CWA	88.0	86.1	26.8	63.0
+ DRAP	89.0	83.2	**30.0**	**63.9**
TI-FGSM	81.9	79.6	18.0	55.8
+ MI	83.7	81.7	18.7	57.2
+ PI	84.6	78.9	18.8	56.9
+ VT	87.1	89.2	18.6	60.2
+ RAP	85.7	74.6	18.5	56.1
+ PGN	**89.9**	**93.0**	20.1	62.7
+ CWA	82.7	70.2	24.8	56.4
+ DRAP	89.5	81.5	**29.4**	**63.5**
Admix	86.9	87.5	18.3	59.6
+ MI	87.8	88.8	19.0	60.5
+ PI	87.2	87.2	19.0	59.9
+ VT	88.9	**93.2**	18.9	62.0
+ RAP	89.3	81.7	18.7	59.1
+ PGN	**89.8**	92.3	20.0	62.5
+ CWA	87.6	81.9	25.9	61.5
+ DRAP	90.0	83.1	**29.0**	**63.9**
SIA	89.0	94.2	18.3	62.0
+ MI	90.5	**96.5**	19.5	63.6
+ PI	90.4	96.3	19.5	63.5
+ VT	89.9	95.2	19.3	63.0
+ RAP	87.7	88.4	19.0	60.4
+ CWA	89.6	88.7	26.2	64.0
+ DRAP	**90.5**	84.5	**28.2**	**64.2**

Algorithm 2: Flat-RAP algorithm

Require: benign data$(\boldsymbol{x}, y)$, perturbation budget γ, surrogate model distributions $\{P_{\mathcal{S}_i}\}_{i=1}^{I}$, number of samples within one distribution n, late start iteration number n_{LS}, inner iteration number T, step size β, α, decay factor μ.
Initialize $\hat{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}, \boldsymbol{m} \leftarrow 0$;
for $j = 0, ..., K-1$ **do**
  for $i = 0, ..., I-1$ **do**
    Sample a surrogate model $\boldsymbol{w}_i$ from $P_{\mathcal{S}_i}$;
  end for
  if $j \geq n_{LS}$ **then**
    # Inner maximization
    Initialize $\boldsymbol{\epsilon} \leftarrow 0$;
    for $t = 0, ..., T-1$ **do**
      Calculate $\boldsymbol{g} = \nabla_{\boldsymbol{\epsilon}} \ell \left(\frac{1}{I} \sum_{i=0}^{I-1} f\left(\hat{\boldsymbol{x}}_j + \boldsymbol{\epsilon}, \boldsymbol{w}_i\right), y \right)$;
      Update $\boldsymbol{\epsilon} = \boldsymbol{\epsilon} + \beta \cdot \operatorname{sign}(\boldsymbol{g})$;
    end for
  end if
  # Outer minimization
  Calculate $\boldsymbol{g} = \nabla_{\hat{\boldsymbol{x}}} \ell \left(\frac{1}{I} \sum_{i=0}^{I-1} f\left(\hat{\boldsymbol{x}}_j + \boldsymbol{\epsilon}, \boldsymbol{w}_i\right), y \right)$;
  Update momentum by $\boldsymbol{m} = \mu \cdot \boldsymbol{m} + \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_1}$;
  Update $\hat{\boldsymbol{x}}_{j+1}$ by $\hat{\boldsymbol{x}}_{j+1} = \Pi_\gamma \left(\hat{\boldsymbol{x}}_j - \alpha \cdot \operatorname{sign}\left(\boldsymbol{m}\right)\right)$;
end for
Return $\hat{\boldsymbol{x}}_K$.

in Table XII, for a fixed number of iterations used to update AE, attacks with more iterations penalizing sharpness can effectively improve the attack performance over $n_{LS} = K$. The peak performance is observed at approximately $n_{LS} = 5$, underscoring the effectiveness of the late-start strategy.

Algorithm 3: Flat-CWA algorithm

Require: benign data(x, y), perturbation budget γ, surrogate model distributions $\{P_{\mathcal{S}_i}\}_{i=1}^{I}$, number of samples within one distribution n, inner iteration number T, step size r, β, α, decay factor μ.
Initialize $\hat{\boldsymbol{x}}_0 \leftarrow \boldsymbol{x}, \boldsymbol{m} \leftarrow 0$;
for $j = 0, ..., K-1$ **do**
  for $i = 0, ..., I-1$ **do**
    Sample a surrogate model $\boldsymbol{w}_i$ from $P_{\mathcal{S}_i}$;
  end for
  # Inner maximization
  Calculate $\boldsymbol{g} = \nabla_{\boldsymbol{x}} \ell \left(\frac{1}{I} \sum_{i=0}^{I-1} f\left(\hat{\boldsymbol{x}}_j, \boldsymbol{w}_i\right), y \right)$;
  Update $\hat{\boldsymbol{x}}_j$ by $\hat{\boldsymbol{x}}_j^0 = \Pi_\gamma \left(\hat{\boldsymbol{x}}_j + r \cdot \operatorname{sign}(\boldsymbol{g})\right)$;
  # Outer minimization
  for $i = 0, ..., I-1$ **do**
    Calculate $g = \nabla_{\boldsymbol{x}} \ell \left(f\left(\hat{\boldsymbol{x}}_j^i, \boldsymbol{w}_i\right), y \right)$;
    Update momentum by $\boldsymbol{m} = \mu \cdot \boldsymbol{m} + \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$;
    Update $\hat{\boldsymbol{x}}_j^{i+1}$ by $\hat{\boldsymbol{x}}_j^{i+1} = \Pi_\gamma \left(\hat{\boldsymbol{x}}_j^i - \beta \cdot \boldsymbol{m}\right)$;
  end for
  Calculate the update $\boldsymbol{g} = \hat{\boldsymbol{x}}_j^I - \hat{\boldsymbol{x}}_j$;
  update $\hat{\boldsymbol{x}}_{j+1}$ by $\hat{\boldsymbol{x}}_{j+1} = \Pi_\gamma \left(\hat{\boldsymbol{x}}_j + \alpha \cdot \operatorname{sign}(\boldsymbol{g})\right)$;
end for
Return $\hat{\boldsymbol{x}}_K$.

B. Choice of Architecture within Prototype

In DRAP, we improve between-distribution diversity by choosing surrogate distributions on model weights of architectures from diverse prototypes. *Does the choice of architecture within a prototype have as significant an impact on improving transferability as the prototypes themselves, as shown in Section VI-C2?* For comparison, we substitute the surrogate model ResNet-50 in default protocol in Section VI-A with VGG-19-BN, Inception-V3 and DenseNet-121 and report the attack success rated averaged on the whole target model sets in Table XIII. We can observe that DRAP is less sensitive to different convnets. Even with other convnets, the attack performance is still quite decent. This is because even if they provide diversity in model distribution, they give rise to similar loss landscape from the adversarial perspective. The results demonstrate that the between-distribution diversity among architectures, a lower-level concept than diversity among prototypes, is less important to transferability.

APPENDIX E
FULL IMAGENET TARGETED RESULTS

In Table XIV, we present targeted attack results on ImageNet dataset broken down to each target models for the methods evaluated in Table III.

TABLE XI
THE NUMBER OF GRADIENT CALCULATIONS REQUIRED IN DIFFERENT METHODS. ASIDE FROM DRAP, ONE UPDATE DIRECTION CALCULATION OF OTHERS REQUIRES I QUERIES OF MODEL IN SURROGATE ENSEMBLE. WE USE THE SAME UNTARGETED EXPERIMENTAL PROTOCOL AS IN SECTION VI-A ($I = 5$). FOR RAP, THE LATE-START K_{LS} IS REDUCED IN PROPORTION TO 50.

Attack	**# of gradient calculations (N_g) vs. n_{iter}**	**Hyper-parameters**	$N_g(n_{iter} = 25)$
MI-FGSM	$n_{iter} \times I$	\	125
PI-FGSM	$n_{iter} \times I$	\	125
RAP	$\begin{cases} n_{iter} \times I, \ n_{iter} < K_{LS} \\ K_{LS} \times I + (n_{iter} - K_{LS}) \times (T+1) \times I, n_{iter} \geq K_{LS} \end{cases}$	$K_{LS} = \begin{cases} 0, n_{iter}/I < K_{LS} \\ 50, n_{iter} \geq K_{LS} \end{cases}, \ T = 10$	1375
CWA	$n_{iter} \times 2I$	\	250
PGN	$n_{iter} \times N \times 2I$	$N = 20$	5000
DRAP	$\begin{cases} n_{iter}, \ n_{iter} < n_{LS} \times I \\ n_{LS} \times I + (n_{iter} - n_{LS} \times I) \times (T+1), n_{iter} \geq n_{LS} \times I \end{cases}$	$n_{LS} = \begin{cases} 0, n_{iter}/I < n_{LS} \\ 5, n_{iter}/I \geq n_{LS} \end{cases}, T = 5$	150

TABLE XII
ATTACK SUCCESS RATES (%) OF DRAP WITH DIFFERENT LATE START ITERATION NUMBER n_{LS}.

n_{LS}	**CNN Set**	**Metaformer Set**	**CNN(AT) Set**	**Metaformer(AT) Set**	***Overall Average***
0	78.8	42.2	24.9	20.3	45.6
5	80.2	42.6	24.9	20.2	46.1
15	80.3	41.9	24.7	20.2	45.9
25	79.8	42.4	25.2	20.1	46.0
35	78.8	40.7	24.9	20.1	45.2
40	77.7	36.3	24.1	20.0	43.6

TABLE XIII
AVERAGE ATTACK SUCCESS RATES (%) OF DRAP WITH DIFFERENT CONVENETS.

Attack	**Architecture for ConvNet**			
	ResNet-50	VGG-19-BN	Inception-V3	DenseNet-121
DRAP	46.2	44.7	44.5	46.1

TABLE XIV
TARGETED ATTACK SUCCESS RATES (%,↑) ON IMAGENET DATASET. **BOLD** DENOTES THE BEST RESULTS AND UNDERLINED DENOTES THE SECOND BEST RESULTS.

Target Model Set		I-FGSM	DI2-FGSM	SI-FGSM	Admix	TI-FGSM	SSA	SIA	MI-FGSM	PI-FGSM	VT-FGSM	PGN	CWA	SVRE	RAP	**DRAP**
ConvNet Set	AlexNet [45]	0.0	0.1	0.0	0.1	0.1	2.1	1.7	0.0	0.2	0.0	3.2	12.9	0.6	0.3	**30.4**
	VGG-16-BN [46]	9.0	38.1	13.7	20.1	13.2	32.9	**93.3**	7.0	13.4	10.9	50.7	21.9	16.3	29.0	80.0
	DenseNet-201 [47]	10.4	48.6	34.1	24.0	18.4	50.5	**94.7**	13.6	24.8	14.8	62.7	53.4	26.4	28.4	90.7
	GoogLeNet [48]	1.1	13.5	9.0	4.1	3.1	19.6	61.4	2.2	5.0	3.1	31.9	36.9	7.3	9.4	**79.3**
	ShuffleNetV2 [49]	0.3	3.9	2.9	1.5	0.7	7.9	27.7	1.7	2.2	0.8	21.3	33.3	3.0	4.5	**75.7**
	MobileNetV2 [50]	3.6	19.8	9.0	10.5	6.2	27.2	77.3	3.5	7.2	4.3	41.6	37.4	12.5	16.4	**86.3**
	MobileNetV3-L [51]	0.9	8.0	4.6	3.6	2.2	36.3	47.2	2.1	3.8	2.3	29.7	40.1	19.7	6.4	**81.7**
	MNASNet [52]	3.0	16.2	7.5	7.7	5.2	20.9	73.8	3.8	6.8	4.8	40.5	38.5	9.4	14.4	**87.1**
	EfficientNet [53]	0.9	7.3	4.3	1.6	1.1	12.7	44.5	0.9	1.8	1.5	21.5	31.9	5.7	3.3	**72.3**
	ConvNeXt-L [54]	24.6	64.6	31.3	39.4	31.8	87.9	**98.7**	17.9	32.1	26.7	70.7	53.2	84.4	33.1	92.0
	Average	5.4	22.0	11.6	11.3	8.2	29.8	62.0	5.3	9.7	6.9	37.4	36.0	18.5	14.5	**77.6**
Metaformer Set	ViT-S [55]	0.1	4.5	0.3	0.4	0.2	10.4	35.2	0.1	0.4	0.1	7.3	24.6	6.2	1.1	**59.4**
	DeiT-S [56]	0.4	6.6	1.2	0.8	0.4	14.9	37.8	0.7	1.1	0.6	13.5	29.9	8.4	1.5	**71.6**
	PoolFormer-S [38]	2.4	41.6	8.3	8.7	6.4	49.0	**89.5**	5.3	8.9	6.1	45.8	34.8	29.9	15.5	82.4
	TNT-S [57]	0.2	8.4	1.2	0.5	0.3	16.3	47.1	0.2	1.1	0.5	14.9	27.5	11.8	2.3	**72.0**
	Swin-S [58]	0.0	4.5	0.3	0.3	0.2	6.5	31.7	0.3	0.3	0.5	4.0	12.9	7.8	1.0	**37.4**
	XCiT-S [59]	0.0	8.8	0.0	0.4	0.1	10.1	26.4	0.2	0.2	0.1	3.8	9.5	4.9	0.5	**30.2**
	CaiT-S [60]	0.1	7.0	0.3	0.4	0.1	8.1	22.6	0.0	0.5	0.0	3.8	13.8	4.3	0.1	**41.5**
	Average	0.5	11.6	1.7	1.6	1.1	16.5	41.5	1.0	1.8	1.1	13.3	21.9	10.5	3.1	**56.4**
ConvNet(AT) Set	RaWideResNet-101-2 [61]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.1	**7.5**
	WideResNet-50-2 [62]	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.1	5.9	0.2	0.2	**12.3**
	ResNet-50 [63]	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	1.3	0.0	0.0	**4.2**
	ConvNeXt-L [64]	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	4.3	0.1	0.2	**8.2**
	ConvNeXt-B [64]	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	4.4	0.0	0.1	**9.7**
	ConvNeXt-L-ConvStem [65]	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	3.7	0.1	0.2	**8.0**
	ConvNeXt-B-ConvStem [65]	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.1	5.0	0.1	0.1	**9.1**
	Inc-v3$_{ens3}$ [66]	0.0	0.1	0.1	0.0	0.0	0.8	0.3	0.0	0.0	0.0	0.2	13.1	0.4	0.0	**26.4**
	Inc-v3$_{ens4}$ [66]	0.0	0.0	0.0	0.0	0.0	0.9	0.2	0.0	0.0	0.0	0.2	13.3	0.4	0.1	**27.0**
	IncRes-v2$_{ens}$ [66]	0.0	0.0	0.0	0.0	0.0	0.3	0.2	0.0	0.0	0.0	0.3	9.2	0.1	0.0	**21.6**
	Average	0.0	0.0	0.0	0.0	0.0	0.3	0.1	0.0	0.1	0.0	0.1	6.4	0.1	0.1	**13.4**
Metaformer(AT) Set	Swin-B [64]	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	3.5	0.1	0.2	**7.0**
	Swin-L [64]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0	0.2	**8.9**
	XCiT-L [67]	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.1	13.1	0.0	0.2	**17.6**
	ViT-B-ConvStem [65]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	9.1	0.0	0.1	**12.1**
	Average	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	7.5	0.0	0.2	**11.4**
Overall Average		1.8	9.7	4.1	4.0	2.9	13.4	29.4	1.9	3.6	2.5	15.1	19.5	8.4	5.4	**43.5**