SoK: Pitfalls in Evaluating Black-Box Attacks

Fnu Suya*
University of Maryland College Park
suya@umd.edu

Anshuman Suri*
University of Virginia
anshuman@virginia.edu

Tingwei Zhang
Cornell University
tz362@cornell.edu

Jingtao Hong
Columbia University
jh4760@columbia.edu

Yuan Tian
University of California Los Angeles
yuant@ucla.edu

David Evans
University of Virginia
evans@virginia.edu

***Abstract*—Numerous works study black-box attacks on image classifiers, where adversaries generate adversarial examples against unknown target models without having access to their internal information. However, these works make different assumptions about the adversary's knowledge, and current literature lacks cohesive organization centered around the threat model. To systematize knowledge in this area, we propose a taxonomy over the threat space spanning the axes of feedback granularity, the access of interactive queries, and the quality and quantity of the auxiliary data available to the attacker. Our new taxonomy provides three key insights. 1) Despite extensive literature, numerous under-explored threat spaces exist, which cannot be trivially solved by adapting techniques from well-explored settings. We demonstrate this by establishing a new state-of-the-art in the less-studied setting of access to top-k confidence scores by adapting techniques from well-explored settings of accessing the complete confidence vector but show how it still falls short of the more restrictive setting that only obtains the prediction label, highlighting the need for more research. 2) Identifying the threat models for different attacks uncovers stronger baselines that challenge prior state-of-the-art claims. We demonstrate this by enhancing an initially weaker baseline (under interactive query access) via surrogate models, effectively overturning claims in the respective paper. 3) Our taxonomy reveals interactions between attacker knowledge that connect well to related areas, such as model inversion and extraction attacks. We discuss how advances in other areas can enable stronger black-box attacks. Finally, we emphasize the need for a more realistic assessment of attack success by factoring in local attack runtime. This approach reveals the potential for certain attacks to achieve notably higher success rates. We also highlight the need to evaluate attacks in diverse and harder settings and underscore the need for better selection criteria when picking the best candidate adversarial examples.**

I. Introduction

Machine learning models, including models using deep learning, are well known to be vulnerable to specially-crafted inputs, known as *adversarial examples* (AEs), that are designed to induce incorrect predictions. Most early studies of adversarial examples focused on white-box settings where the adversary has full access to the target model [1, 2]. Black-box settings consider scenarios where the adversary has limited access to the target model. Such settings are a more practical threat to many deployed systems [3–5] where the model is not revealed directly. In these attacks, known as *black-box* or *API-only* attacks, the adversary can interact with the target model using API queries but does not have direct access to the model's parameters and may have varying degrees of knowledge about the model architecture, training data, and training process. Previous surveys of such attacks [6, 7] categorize representative attacks based on their adopted methods but overlook differences in assumptions about the adversary's knowledge and capabilities. These assumptions can vary wildly, depending on the resources available for the attacker and the kind of access to the model the API provides. Different assumpations have a significant impact on what attacks are possible in practice. Furthermore, attack evaluations typically rely solely on attack success rates (and query cost for interactive attacks), ignoring how attack success varies across different examples and tasks. This disconnect makes it hard to map out the threat space, leading to improper evaluation of attacks and limiting our understanding of the actual threats.

Contributions. We started by surveying black-box attacks on image classifiers published in major security (Usenix Security, IEEE S&P, CCS, NDSS), machine-learning (ICML, NeurIPS, ICLR, KDD, AAAI, IJCAI) and computer vision (CVPR, ICCV, ECCV) venues. In particular, we identified relevant papers published in the aforementioned top-tier conferences by searching with keywords "transfer", "attack", "black-box", and "query" from the year 2014 (the year of the first paper [1] on generating adversarial examples on deep neural networks) to 2023. In addition to these works, we conducted a thorough search of papers referenced within them and of relevant works citing these identified papers, covering both peer-reviewed papers and preprints online with the best effort. This leaves us with 164 attacks, of which 102 are published in major security and machine learning conferences. With the surveyed attacks, we propose a new taxonomy for existing black-box attacks, organized around assumptions on their threat models. Our taxonomy spans four dimensions (Section IV): 1) interactive queries to the target model allowed, 2) information provided by the target model's API, 3) quality of the initial auxiliary data available to the adversary, and 4) quantity of the initial auxiliary data available to the adversary. These dimensions are chosen based on the underlying components that enable successful black-box attacks in practice—the feedback

*Both authors contributed equally.

available for the attackers to adjust the strategy (whether interactive queries are permitted, and the granularity of the feedback provided if any) and the resources attackers can leverage (quantity and quality of data initially available for the attacker, as well as the availability of pretrained models online, independent from the auxiliary data) We categorize the existing literature using our proposed taxonomy (Section V), focusing on image classifiers as the most widely studied domain. Our observations result in three key findings:

1) Most prior works are concentrated in specific regions of the taxonomy, with several important and practically relevant settings that have not been well explored. Much of this knowledge gap is also likely a technical gap, and we demonstrate this with preliminary experiments on devising stronger baselines in one of the under-explored settings. Despite establishing state-of-the-art attack success, many methods fall short of attacks from more restrictive but well-explored settings, reinforcing the importance of investing research into these under-explored areas (Section VI-A).
2) Some works propose new attacks and compare them to existing baseline attacks under threat models more restrictive than their own, which can underestimate the potency of baselines given enough knowledge. We empirically demonstrate how claims of methods outperforming previous ones can often be invalidated when prior attacks are adapted to and compared under the same threat model (Section VI-B).
3) A closer look at the threat space reveals the scope for utilizing available resources in different and potentially better ways. In particular, attackers with access to some initial auxiliary data and pre-trained models may leverage advances in model extraction [8, 9] and model inversion attacks [10, 11] to enable stronger attacks. We discuss the possible usage of this interaction and motivate future research along this direction (Section VI-C).

Transfer attack evaluations in the literature focus on the number of local optimization iterations as a normalizing factor when comparing attacks. While well intended, such measures are misaligned with practical adversaries' goals: picking an attack that maximizes success within some given time frame. Our evaluation of transfer attacks 1) shows how normalizing for time allows some attacks to run for more iterations and achieve higher success rates; 2) motivates future research to work on better metrics to select better local candidates of adversarial examples, and to evaluate attacks in diverse and harder attack settings. We clarify that adversaries may conduct training [12, 13] with prediction-time attacks. While such adversaries can be extremely potent, our current taxonomy focuses on prediction-time attacks and thus does not capture dynamically changing, possibly poisoned, target models [14].

To support comprehensive evaluations of attacks and defenses, we provide a modular codebase at https://github.com/iamgroot42/blackboxsok

II. Background

We first introduce background on adversarial examples (Section II-A), and then review related works (Section II-B).

A. Introduction of Adversarial Examples

In image classification tasks, given a model/classifier f that takes input $\boldsymbol{x}$ (with ground truth label $c(\boldsymbol{x})$) and generates a prediction $f(\boldsymbol{x})$, the goal of adversary is to achieve some attack goals by adding an (imperceptible) bounded perturbation $\boldsymbol{\delta}$ onto $\boldsymbol{x}$. Depending on the attack goals, there can be *untargeted* and *targeted* attack goals. Untargeted attacks aim to induce a predicted class on the perturbed input $\boldsymbol{x}+\boldsymbol{\delta}$ that is different from $c(\boldsymbol{x})$, namely, $f(\boldsymbol{x}+\boldsymbol{\delta}) \neq c(\boldsymbol{x})$. Note that we define the attack goal of misclassification with respect to the ground-truth label $c(\boldsymbol{x})$ of input $\boldsymbol{x}$, which is consistent with implicit assumptions made in the surveyed black-box attacks (i.e., the evaluations consider misclassifying correctly labeled samples and assume $c(\boldsymbol{x}) = c(\boldsymbol{x}+\boldsymbol{\delta})$). However, there can be other definitions of untargeted attacks that are more related to the definition of the adversarial risk of a model f, such as causing misclassification with respect to $f(\boldsymbol{x})$ or $c(\boldsymbol{x}+\boldsymbol{\delta})$ (if different from $c(\boldsymbol{x})$). Diochnos et al. [15] provide a more detailed comparison between these definitions, but they are the same for the setting considered in this paper. Targeted attacks ensure the perturbed sample $\boldsymbol{x}+\boldsymbol{\delta}$ is misclassified into a particular label $\hat{y}$ that is in the interest of the adversary, namely, $f(\boldsymbol{x}+\boldsymbol{\delta}) = \hat{y}$. The bounded perturbation is $\boldsymbol{\delta}$ is constrained by some perturbation budget ϵ to avoid raising suspicion, although some works also consider minimizing the perturbation magnitude [16, 17]. The most common constraint is to limit the ℓ_p norm of the perturbation $\boldsymbol{\delta}$, namely $\|\boldsymbol{\delta}\|_p \leq \epsilon$.

The white-box attacks have access to all the internal information of the target model and therefore, can optimize the perturbation $\boldsymbol{\delta}$ with respect to (w.r.t) some loss function (e.g. maximize cross-entropy w.r.t $c(\boldsymbol{x})$ in untargeted and maximize the loss w.r.t $\hat{y}$ in targeted settings) to generate the adversarial examples using gradient descent [2]. In contrast, black-box attacks do not have access to the model's internal information and, therefore, either rely on transfer attacks if some local surrogates are available (Section III-A) or black-box optimization if interactive access is permitted (Section IV-A).

B. Related Works

Surveys on Black-box Attacks. Two survey papers already cover black-box attacks in the vision domain [6, 7]. These papers categorize representative attacks by methods, identifying the best attacks and offering meta-analyses of their reported results. However, they draw conclusions from experimental results reported in prior works, which are spread across incompatible settings and threat models. In contrast to these works, we provide the taxonomy based on the threat model, which enables a better understanding of how attacks relate and how they should be compared. This, in turn, allows us to evaluate attacks in consistent test environments and draw meaningful conclusions. The most relevant previous work is Zhao et al.'s

comprehensive evaluation of transfer attacks in the image domain [18]. They focus on understanding the robustness of different defenses against untargeted transfer attacks at a fixed perturbation budget and compare the visual stealthiness of different attacks with the same norm constraint. In contrast, we focus on general black-box attacks and compare attacks across various threat models.

Relevant SoKs. There are several previous SoK papers on adversarial machine learning, focusing on different topics ranging from categorizing attacks on audio recognition systems [19] to certified robustness for adversarial examples [20]. Papernot et al. provide a general systematization of adversarial machine learning, but do not focus on black-box attacks [21]. Carlini et al. [22] provide a set of guidelines for proper evaluation of adversarial robustness in white-box settings. While some recommendations, such as proper threat model categorization and running attacks until convergence, apply to black-box attacks as well, we provide a concrete taxonomy with detailed analyses and advocate for time-based comparison of attacks. Our paper is the first to systematize knowledge of black-box attacks based on their applicable threat models.

III. Attack Methods

Most attacks use the same underlying principles and structure, but make advances in one or more aspects of the attack process. In this section, we categorize attacks based on their strategies, building on top of the categories provided in the prior literature [7, 18]. These categories help better understand similarities and connections between existing attacks and identify scope for improvement and combinations of advancements. This categorization is orthogonal to the threat-model based taxonomy we introduce in Section IV.

A. Transfer attacks

Transfer attacks first generate adversarial examples for local surrogate models with white-box access and then attempt to transfer those local adversarial examples to the target model [2, 23]. The success of a transfer attack depends on how similar (at least with respect to the relevant decision boundaries) the local models are to the target model and how effective the local attack is at finding generalizable adversarial examples against the local models. We mainly adopt the terminology used by Zhao et al. [18] to describe existing attacks. However, we added a category of *Better Loss Functions*, which customizes the loss function for better transferability.

Gradient Stabilization. The idea behind the gradient stabilization is to make the model less prone to overfitting to the local model and improve the transferability to the unknown target model, through utilizing the spatial [24–32] and temporal correlation [30, 31, 33–40] among the gradients.

Input Augmentation. The main idea is to augment and diversify inputs to increase the transferability of adversarial examples to different target models, and is similar to generating more generalizable models through augmenting the training data [18, 41]. Early works applied common and hand-crafted input augmentations [32, 34, 37, 42–48]. Recent works have focused on learning better input transformations with neural networks [49] or finding the best input from existing ones using neural networks [50, 51] or reinforcement learning [52].

Better Loss Functions. Recently, more research is focused on leveraging intermediate model features [31, 53–67] instead of the model outputs, with an intuition that intermediate feature representations may be more generic and transferable [68, 69], which can be further enhanced with interpretability techniques to focus on relevant features [70, 71].

Surrogate Refinement. Different local surrogate models can have different transferability to the target model, and the most naivë approach to improve transferability is to adopt an ensemble of models [72, 73]. Other techniques focus on improving transferability of a single surrogate, such as modifying for or using better local architectures (e.g., using skip connections) [74–76]; modifying the activations functions [77–80]; modifying the training strategy (e.g., adopting adversarial training, early stopping) [81–87]; identifying proper source models with meta-learning [32, 88].

Generative Models. Unlike iterative attacks, existing works also train generative models that, given an input, produce the corresponding adversarial example. Existing works focus on using better loss functions to train the generators [89–92] and also using better generator architectures such as class-specific [93] or class conditional ones [87].

B. Query-based attacks

Query-based attacks refine the candidate adversarial examples with interactive queries until the attacker's objective is achieved. Below, we introduce the common methodologies which are based on gradient-estimation or gradient-free attacks. Gradient-free attacks can apply to much broader settings, especially for non-differentiable models, whereas gradient-estimation methods are illogical for such models. While estimating a non-differential target model with certain differentiable approximations is possible, this remains an open question. For deep neural networks, gradient estimation methods perform better on tasks that involve minimizing the perturbation magnitude [94, 95], while gradient-free attacks tend to work better under fixed perturbation budgets, especially for the ℓ_∞-norm [96, 97].

1) Gradient-estimation Attacks: These attacks work by estimating the gradients of the unknown target model and updating candidate adversarial examples accordingly. This technique can be applied irrespective of the target models returning full prediction scores [3, 16] or just the prediction class [98–100].

Complete Confidence Vector. In this setting, confidence scores of all classes are available, and attacks start from the original seed image and gradually search for better perturbations with the estimated gradients. Ever since the first work on estimating the gradient for every coordinate with finite-difference method [16], subsequent works focused on finding more efficient gradient estimation strategies, mostly by finding

a better random perturbation vector to estimate the gradient efficiently for the finite-difference method [3, 4, 101–111].

Hard-Label. The hard-label attacks are more restrictive and can only access the prediction label of the highest confident class; hence, the attacks usually require a reference image that satisfies the attacker's objective (e.g., the reference image is from the intended target class for misclassification) to generate a likely-to-succeed perturbation and then focus on minimizing the size of the perturbation (measured by ℓ_p-norm such as ℓ_2) with the estimated gradients. Since the first work [98], various techniques are proposed to improve the gradient estimation quality and boost attack performance [99, 100, 112–116].

2) Gradient-free Attacks: s the name suggests, gradient-free attacks do not rely on estimating the target model gradients. These attacks are diverse in terms of their methodologies.

Complete Confidence Vector. Gradient-free attacks with complete confidence vector range from classical black-box optimization techniques (e.g., genetic algorithms, evolution strategies, Bayesian Optimization) [117–121] to efficient random search strategies [96, 122–128]. The key is to find an effective low-dimensional subspace to generate perturbations and then map back to the original input space. The recent efficient random search-based attacks [96, 124] are the current state-of-the-art to generate norm-bounded perturbations.

Hard-Label. The first type of gradient-free methods are based on random walk with various sampling distributions [129–134] or directions based on the geometry of the decision boundary [135]. Recently, more efficient attacks are proposed using diverse techniques such as random search [97, 136, 137], evolution strategies [138] or utilization of geometric properties of the boundary [139]. For norm-constrained adversaries, especially in ℓ_∞-norm, the random search-based methods achieve the state-of-the-art performance [97, 136].

C. Hybrid Attacks

These attacks utilize surrogate models, like transfer attacks, and submit queries to the target model. We name these attacks "hybrid attacks" to distinguish them from pure transfer or query-based attacks. There are mainly two types of hybrid attacks. The first type leverages surrogate models to enhance query-based attacks by providing better starting points (i.e., warm starting) [140] or providing better sampling space of perturbation [141–147] for the query-based attacks. The second type improves available *surrogate models* with labeled queries from the target model, including fine-tuning the models [140, 143, 148, 149] or finding proper weights for individual models in the model ensemble [150], so that the transferability from these similar models can be significantly improved in the later stage. The only exception from above is that queries from the target model can also be combined with local explanation techniques [151] to select the most transferable single model from a set of classifiers [152].

IV. Taxonomy Threat Model

We propose a new attack taxonomy organized around the threat model assumptions of an attack, using four separate dimensions to categorize assumptions made by each attack. Within each dimension, we describe different categories in order of increasing knowledge available to the adversary (Section IV-A - Section IV-D). We then discuss the existence of pretrained models as a sub-axis Section IV-E and how it may interact with the main axes of our dimension. We then use our taxonomy to categorize attacks (Section V) and report our insights with directions for future research (Section VI).

A. Query Access

Query access captures the adversary's ability to query the target model *before* sending its final adversarial input. We group access levels into two characteristic settings:

(a) **No Interactive Access:** the adversary has absolutely no opportunity to query the target model interactively. Likely scenarios include situations where the adversary has only one-way communication with the target model through an indirect victim. For example, the adversary may want to generate malware that bypasses the victim's malware classification system but without any way to query that system directly. This is the most challenging attack setting where the adversary has no opportunity to learn from feedback from the target model.

(b) **With Interactive Access:** a more relaxed setting and still has wide applications in practice. In this setting, the attacker can interactively query the target model and adjust subsequent queries by leveraging its history of queries. However, the number of queries that can be submitted might be constrained significantly in practical cases,e.g., rate limits imposed by the target model API, the financial cost involved in making queries, or simply the attackers wanting to avoid raising suspicion. In other situations, the attackers may still be able to query the target model as often as they wish. The most concrete example of unlimited black-box query access would be one where the adversary has access to the model on their hardware, but it is encrypted in a secure enclave (e.g., Intel SGX as the Trusted Execution Environment) that protects its parameters [153, 154].

B. API Feedback

This dimension captures the granularity of information the target model's API returns for a given query. We break this down into three distinct categories:

(a) **Hard-Label:** the only value returned by the API is the predicted label for the given query input. For instance, a face-recognition based utility may only provide a label for match/mismatch.

(b) **Top-K:** the model API returns confidence scores for the top-k ($1 \leq k < N$, for N classes) labels. This aligns well with most real-world predictive APIs, which often return confidence values for a few most likely classes to minimize network overhead. This setting provides more

information than hard-label access even when $k = 1$, since the confidence score for the predicted label is made available. For example, Google's Cloud Vision API[1] uses labels from their Knowledge Graph API[2], which has tens of thousands of labels, and returning classification scores for all classes is unlikely to be helpful for benign users.

(c) **Complete Confidence Vector:** the API returns confidence scores for all classes. This may correspond to the enclave-based setting described above, or one where the number of classes is low enough for an API to return all related information.

Below, we describe auxiliary information available to attackers for more efficient attacks. We define two axes of 1) the *quality* of data and 2) the *quantity* of data.

C. Quality of Initial Auxiliary Data

This dimension captures the correctness of the adversary's priors on the target model's training data. Higher quality of auxiliary data indicates that the attackers can conduct the attack without considering potential distributional gaps. In this paper, we capture such distributional gaps using the overlap between the feature or label space of two distributions (corresponding to the target model's data and auxiliary data). Feature space overlap refers to same/similar samples in the data feature (e.g., images of dogs in two distributions) regardless of the assigned labels (e.g., different labels for the same image, depending on different tasks). We discuss overlap on distributions, not on datasets, because distributions are more fundamental than the (sampled) datasets.

(a) **No Overlap:** auxiliary data available to the adversary does not overlap in the data features and the labels. This setting is closest to real-world APIs, where knowledge about the target model's training data is obfuscated and often proprietary (like GPT-4).

(b) **Partial Overlap:** auxiliary data available to the adversary has partial overlaps (in the distributional sense) with the private training data of the target model regarding data features or labels. This setting best matches scenarios where the training data of the target model includes some publicly available datasets.

(c) **Complete Overlap** auxiliary data available to the attacker is the same as the target model's training data, or sampled from the same underlying distribution (i.e., same label space and feature space). For example, the target model could be trained on a publicly available dataset, and this information may be public.

Notably, removing the high overlap in data distributions can significantly undermine the attack success [155]. The authors propose a variant of PGD (masked PGD) to mitigate the performance degradation due to distributional gap.

D. Quantity of Initial Auxiliary Data

Finally, we consider the quantity of auxiliary data (independent of data quality) *initially* available to the adversary. We explicitly mention the availability of initial auxiliary data because the existence of some pretrained models may change the amount of auxiliary data available for the adversary in the end (Section IV-E). We consider two categories: the first is when the amount of data is only a handful and hence cannot be used to train models with decent performance for the attacks, while the second entails situations with enough data to train performant models. Note that the definition of useful performance can vary depending on application scenarios, and we use this hypothetical and abstract description here. In practice, attackers may check whether the amount of available data can be used for training more useful models from the perspective of attack effectiveness (e.g., the threshold can be set as the quantity sufficient to train a surrogate classifier that is only X% off compared to the prediction accuracy of the target model).

(a) **Not Sufficient:** the quantity of data available is insufficient to train models useful for attacks. Attackers in this category may opt for leveraging other ways to utilize this information (e.g., computing sample statistics [175] or training shallow models [168]). This category also contains the scenario of no auxiliary data (i.e., no samples). Strictly speaking, the "quality" of the datasets does not matter as there is no auxiliary data at all, and this category falls ambiguously into any category of "Auxiliary Data Quality". However, for clarity in presentation, we move attacks that do not require any auxiliary data into the category corresponding to the quality of "No Overlap", to (best) denote that these attacks do not require any knowledge from the auxiliary data.

(b) **Sufficient:** the quantity of data available is sufficient to train decent models (e.g., generative models or classifiers), that can in turn assist with attacks.

While attack strategies that require auxiliary data can technically be applied for any amount of data, implicit assumptions in such attacks may dictate certain requirements on data quantity for them to be effective. A discussion around the initial "quantity" of data is thus still relevant. For example, methods that require data to train well-performing surrogate models would understandably suffer from significant performance degradation when the amount of auxiliary data is limited, as demonstrated in ablation studies [91]. However, the paper does not explicitly report the point at which attack performance drops to near-random. On the other hand, methods in "Not Sufficient" categories might face a bottleneck when given sufficient data, as the proposed approaches implicitly assume *limited* data. Ablation studies on the impact of quantity of auxiliary data can be helpful to the community but are currently lacking in the literature. We advocate for including such studies in future works and discuss more in Section VI-B.

E. Existence of Pretrained Models

The literature has been historically building surrogate models directly from target models [172], and the availability of pretrained models today is an artifact of orthogonal advances in machine learning for building and releasing high-performing

[1] https://cloud.google.com/vision/docs/labels

[2] https://developers.google.com/knowledge-graph/reference/rest/v1/

Quality	Quantity	No Interactive Access	With Interactive Access		
			Hard-Label	Top-K	Complete Confidence Vector
None	Insufficient	Frequency Manipulation [156] **w/ Pretrained Surrogate*:** Better Loss: [90–92, 155, 157–165] Better Loss for AE Generator: [90, 91, 162]	Random walk: [129–135] Gradient estimation: [98–100, 112–116] Other Gradient-free: [97, 136–139] Classic Black-box Opt.: [108, 166]	NES [3]	Gradient Estimation: [3, 4, 16, 101–111] Classic Black-box Opt.: [117–121] Efficient Random Search: [96, 117–119, 122–128]
	Sufficient	∅	∅	∅	∅
Partial	Insufficient	**w/ Pretrained Surrogate*:** Better Loss: [92, 155, 158, 163]	∅	∅	Boost Existing Methods w/ Trained Generator: [167]
	Sufficient	∅	∅	∅	∅
Complete	Insufficient	Train Shallow Surrogate: [168, 169] **w/ Pretrained Surrogate*:** (Basic) Gradient Sign: [2, 23] Input Augmentation: [32, 34, 37, 42–52, 170] Gradient Stabilization: [24–40] Better Loss: [31, 53–67, 165] Refine Surrogate: [32, 72–80, 84, 88]	Improve UAP w/ Feedback: [164] Train Surrogate w/ Synthetic Data: [171–174] Boost Existing Methods w/ Unlabeled Data [175]	∅	Boost Existing Methods: Trained Generator: [167, 176–179], Unlabeled Data [175] **w/ Pretrained Surrogate***: Save Queries with Surrogate: [140–149, 151] Refine Surrogate with Queries: [143, 150, 152]
	Sufficient	Train Better (Deep) Surrogate: [81–83, 85, 86] Train AE Generator: [89, 91, 93, 180–182] Input Transformation Network: [49, 50, 52] Train Simple Auxiliary Classifier: [58, 59, 91]	Improved Gradient Estimation w/ Trained Generator: [94, 95]	∅	Train AE Generator: [87, 183–185]

TABLE I: Threat model taxonomy of black-box attacks. The first two columns correspond to the quality and quantity of the auxiliary data available to the attacker initially. The remaining columns distinguish threat models based on the type of access they have to the target model, and for adversaries who can submit queries to the target model, the information they receive from the API in response. The symbol ∅ above corresponds to areas in the threat-space that, to the best of our knowledge, are not considered by any attacks in the literature. The sub-category of *w/ Pretrained Surrogate* with "*" denotes that the corresponding attacks do not require auxiliary data, but the quality of data used to train the surrogate determines the corresponding cell.

models, especially in the image domain. Such an assumption may not hold across other domains, especially in security-critical areas. We refer to such models as "pretrained" models. Assume a pretrained model trained on unknown proprietary data that is highly similar to the target model's data. An adversary that uses such a model implicitly leverages this data overlap through the publicly released model. To better capture this implicit leverage of high data quality, we classify attacks that only involve pretrained models into settings where the quantity of *initial* auxiliary data is zero, and the quality of data is determined by the quality of private data used to train the model. For clarity, we add the existence of pretrained models as a sub-axis on top of the four main axes mentioned above.

V. Classification of Attacks on Threat Model

In this section, we categorize the black-box attacks based on their presumed threat model. Table I presents our categorization of the surveyed attacks. The first main division is between attacks where the adversary has no interactive access to the target model, and ones where some level of interaction is available. Within each of these, we consider threat models based on the quality and quantity of data available to the adversary. For the rest of this paper, we interchangeably use 'transfer attacks' with 'non-interactive attacks', and 'query-based attacks' with 'interactive attacks'.

A. No Interactive Access to Target Model

A significant fraction of attacks in the literature assume an adversary with no ability to submit queries and obtain responses from the target model. Without such access, the adversary has limited options and must use local resources to find good candidate examples.

1) Low Quality Data: No Overlap with Target: This threat model assumes the least adversarial knowledge as the auxiliary data available for the attacker has no overlap with the training data for the target model, and the availability of the auxiliary data is limited. Works in this threat model have only appeared recently and to our knowledge, there is only one work that does not consider additional information (e.g., pretrained models) and obtains successful adversarial examples with frequency manipulation [156]. A relaxation of this setting allows the adversary access to pretrained model(s) where the training set does not overlap with the target. As noted in Table I, attacks in the literature that assume access to a pretrained surrogate do not leverage any additional auxiliary data and therefore, the quantity of auxiliary data is actually zero. Despite having a distribution mismatch, surrogate model(s) can capture some level of image semantics that can be valuable for adversaries. Customized loss functions with respect to the pretrained models are designed by the adversaries to generate successful adversarial examples [90–92, 157–165]. We note that some works [92, 155, 158, 163] relax their setting to allow the auxiliary data to have partial overlap with the target in the data points and/or the labels, and also the availability of pretrained models (trained on data with partial overlap with target). These attacks still design customized loss functions to cope with distribution mismatch, and the (minor) difference to

the "no" overlap setting mainly lies in how to map the labels of the local surrogate to the labels of the target. As expected, attacks in partial-overlap settings achieve better results than ones with no overlap. To the best of out knowledge, no work in the literature assumes sufficient low-quality (no/partial overlap with target) auxiliary data, while this situation is likely to be common in practice. For example, when attacking some unknown target model (e.g., medical image classifier [186]), attackers may leverage the ImageNet dataset.

2) High Quality Data: Complete Overlap with Target: The distribution of auxiliary data is highly similar (or even the same) to the target training distribution. Under limited availability of such data, shallow surrogates can be trained to enable higher transferability [168, 169]. This assumption may be further relaxed when adversaries have access to some pretrained models trained on high-quality auxiliary data. Like the case of low-quality auxiliary data, existing works that use pretrained models do not utilize auxiliary data. This is the most explored attack setting in the literature: methods include gradient stabilization [24–40], input augmentation [32, 34, 37, 42–52, 170], better loss designs [31, 53–67, 165] and surrogate refinement [32, 72–75, 77–79, 81–88], as discussed in Section III-A. One example of a scenario with an insufficient amount of high-quality auxiliary data is the case of a face recognition target model. In this context, auxiliary data might only consist of a few face images captured under the same conditions (such as the same setting, background, etc.) as the target model's training data, but acquiring a large amount of such high-quality data can be challenging.

When there are sufficient amount of high-quality auxiliary data available (attackers can also naturally obtain well-performing surrogate models), the proposed methods can be quite diverse: directly training better surrogate classifiers to generate more transferable adversarial examples [81–83, 85, 86], training auxiliary classifiers on top of the surrogate classifiers [58, 59, 91], training generators to generate likely-to-transfer adversarial examples [89, 91, 93, 180–182]. Besides these methods, some attacks also focus on finding better transformation methods with neural networks [49, 50, 52] so that these inputs, when input to some surrogate classifiers, can lead to improved transferability. Notably, many of these attacks (e.g., training auxiliary classifiers and finding better input transformations) are compatible with each other, indicating that stronger attacks might be possible by composing these attacks, which are not explored in the literature, and we encourage researchers to investigate this possibility.

B. Hard-Label with Interactive Access

In this subsection, we consider attacks where the adversary can actively query the target model, but only receives hard-label responses. Within this category, we break down attacks according to the auxiliary data available to the adversary, following a structure similar to that of the previous subsection.

1) Low Quality Data: No Overlap with Target: Attacks in this category are rather restricted in terms of the attacker knowledge as existing attacks in the literature in fact did not utilize any auxiliary data, leading to the category of the quantity of auxiliary data being zero. The quality of data should not be relevant in this case, but we still put it into the setting of "no overlap with target" mainly for convenience in categorization. Despite being a challenging setting, many hard-label query-based attacks are proposed. The common methods include estimating the gradients [98–100, 112–116], deploying some classic black-box optimization techniques [108, 166], leveraging random-walk strategy [129–135], or developing other gradient-free random search based methods [97, 136–139]. The categories that allow adversaries to leverage some (sufficient or insufficient but not zero) amount of auxiliary data are largely missing from the current literature.

C. High Quality Data: Complete Overlap with Target

Attacks in this category have access to auxiliary data sampled from a distribution highly similar to/same as the target distribution. When the amount of auxiliary data is insufficient, the proposed methods include finding better (untargeted) universal adversarial perturbations that are agnostic to the victim images [164], training surrogate models using synthetic dataset [171–174] and boost existing hard-label attacks using limited amounts of unlabeled dataset [175]. When sufficient auxiliary data is available, this data can be used to train generators to obtain better gradient estimates [94, 95]. Interestingly, the number of works published under this category is still much less compared to the more restrictive category above.

D. Top-K Confidence Vector with Interactive Access

Attacks in this category can interact with the target model and get the top-k part of the confidence vector from the target model. So far, there is only one work [3] that explicitly designs an attack for this setting, although such a scenario is also very common in practice. Driven by limited exploration in this category, we conduct preliminary experiments in Section VI-A to show that, currently under-explored areas may not be solved by trivially adapting techniques from other well-explored areas and motivate future investigation along this direction.

E. Complete Confidence Vector with Interactive Access

Attacks in this category will receive the complete prediction confidence vector returned from the target model. The remaining breakdowns are still similarly based on the quality and quantity of the auxiliary data available.

1) Low Quality Data: No Overlap with Target: The strictest setting is when the adversaries do not use any auxiliary data. In this setting, many works propose generating highly successful adversarial examples (e.g., finding many untargeted adversarial examples in < 100 queries). Typical methods include gradient estimation [3, 4, 16, 101–111], leveraging classical black-box optimization techniques [117–121] or proposing some efficient random search methods [96, 117–119, 122–128].

When the assumption is relaxed to allow limited number of auxiliary data that overlaps with the target distribution partially, a generator [167] on the perturbation distribution can be trained to boost the performance of the state-of-the-art Square Attack [96] that does not use any auxiliary data.

2) High Quality Data: Complete Overlap with Target: Under limited availability of high-quality auxiliary data, existing works train generators to improve performance by better capturing the low-dimensional latent space where the adversarial examples reside [167, 176–179]. The availability of some (auxiliary) unlabeled data also improves existing attacks that (originally) do not rely on auxiliary data [175]. Like the low-quality data case, a generator for the perturbation distribution can still be trained on the limited high-quality auxiliary data to boost performance of the Square Attack [167]. When the assumption is further relaxed to allow some pretrained models trained on data highly similar to the target's training data, the pretrained models can be used to boost query-based attacks [140–149, 151] or queries from the target model can be used to refine the surrogate model [143, 150, 152]. The most relaxed setting is when there are sufficient high-quality auxiliary data available. Existing works train generators on (sufficient) high-quality data to generate adversarial examples directly. In particular, a generator is first trained on some local surrogate models (can be easily obtained by training on the auxiliary data if not available beforehand) and later fine-tuned with queries from the unknown target model [87, 183–185].

VI. Insights from Taxonomy

Studying published attacks from the perspective of our threat model taxonomy results in several insights about gaps in the current research (Section VI-A), ways to improve evaluation (Section VI-B), and opportunities to improve techniques by incorporating ideas from related fields, such as model extraction and inversion (Section VI-C).

A. Technical challenges in Underexplored Areas

As can be seen from Table I, many threat models are unexplored (marked with ∅) or have only been considered by a few works. Across the rows, there is little work in settings where ample data is available but from sources that have limited overlap with the target model's data distribution. However, this is perhaps the most relevant practical scenario—for most classification tasks, adversaries are likely to be able to acquire large amounts of somewhat similar data (e.g., from the Internet, open image datasets), but unlikely to be able to sample from the same distribution as the target model's (private) training distribution.

Across the columns, only one attack explicitly optimizes for the availability of top-k prediction scores. This is surprising since this is the most likely scenario for API attacks on deployed classifiers. For example, ClarifAI's models[3] return scores for at most 200 classes. For these unexplored or under-explored settings, we suspect there is a technical gap in addition to a knowledge gap, so the settings cannot be addressed satisfactorily by adapting state-of-the-art methods from well-explored areas [92, 155]. To support our argument, we propose an attack for the top-k setting, specifically for the setting with no auxiliary data or pretrained models is available, the typical setting for query-based attacks (Section III-B). Our adapted attack is based on the Square Attack [96] that is originally designed for the setting that receives full confidence vector of prediction and the adaptation idea is built on top of the design of NES: top-k attack in Ilyas et al. [3] with non-trivial modifications (details in Appendix A3).

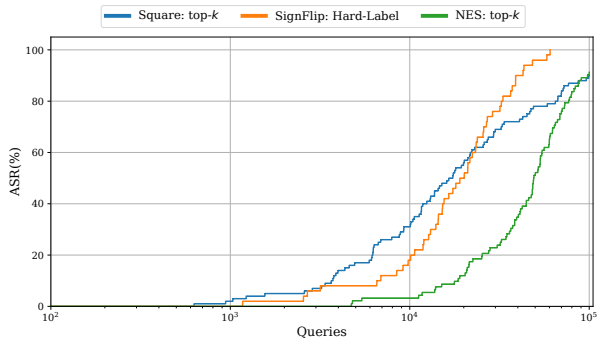

Fig. 1: Comparison of top-k attacks. Square: top-k is our proposed adaption of the Square Attack for the top-k setting. NES: top-k is the current state-of-the-art attack. SignFlip [136] is a more restrictive hard-label attack.

Following the setup in the baseline NES: top-k [3], we consider targeted attacks, set the query limit to 100,000 for both attacks, and assume only the top-1 prediction confidence is available. As shown in Figure 1, the Square: top-k outperforms the NES: top-k attack significantly as the number of queries increases. However, the Square: top-k is mostly outperformed by the hard-label SignFlipAttack [136], which ensures the targeted label y_t is always in the top-1 prediction and then chooses to ignore the extra prediction confidence. This comparison illustrates that there is substantial room for improving attacks in the top-k setting, as attacks designed for this setting are not performing nearly as well as attacks with less information. Moreover, it is essential to underscore the significance of considering baseline attacks that operate with a subset of the available information in the given context. One might improve further top-k attack performance by harnessing the available confidence score for hard-label attacks or adapting techniques from multi-label learning [187]. However, this task is not straightforward, as our preliminary experiments suggest. In general, research in the underexplored areas, such as the ones we outline, faces two unique challenges. First, well-explored methods that require extra information do not directly extend to other threat models, and their adaptation can be complex, with no reasonable estimates on the performance drops—we demonstrate one such case in our experiments in Figure 1. Second, well-explored attacks from more restrictive settings (i.e., less information) can be trivially extended to less restrictive settings, but the room for improvement with additional information provided by such less restrictive settings is unclear. For instance, the state-of-the-art hard-label attack has a success rate of 20% at 10,000 query limit, while the full-score setting has near-perfect attack success at the same number of queries. While one can

[3]https://docs.clarifai.com/api-guide/predict/prediction-parameters/

Attacks	Square Attack	ODS-RGF	Hybrid-Square
Attack Success (%)	100	97.7	100
Average Queries	2,317	1,242	117

TABLE II: Comparing query-based attacks in a setting where all attacks are given access to an ensemble of four surrogate models. Experimental setup follows from Tashiro et al. [142]. Hybrid-Square is our proposed stronger baseline.

argue that hard-label attacks can be used in top-k setting and yield 20% success rate, it is intuitively clear that additional information via top-k should be usable to increase success rates significantly. While our adaptation (Figure 1) achieves a success rate of over 30%, there may still be room for improvement. We thus encourage researchers to concentrate on crafting and examining attacks designed for relevant threat model scenarios, such as the ones we identify.

B. Stronger Baselines Under Same Threat Model

Works introducing black-box attacks often make various assumptions about the knowledge of the adversary and often end up comparing adversaries across different levels of knowledge directly in terms of attack effectiveness. We advocate that with the categorization of the threat space (as outlined in Table I), attacks should be carefully compared within the same threat space. Further, researchers should be mindful of the possibility of combining additional information made available to the adversary to design stronger baselines.

Here, we use a preliminary experiment on the category of complete access to prediction vectors and an ensemble of local surrogates to demonstrate that, when evaluated under the same threat model, a strong baseline can exist (and be easily found) to overturn the state-of-the-art claims in the paper. Specifically, ODS-RGF [142] leverages diversified gradient vectors from the local surrogate models as the perturbation vector for the RGF attack [145]. This attack performs better than the Square Attack [96] that does not require any pretrained surrogate models. Using a simple strategy of generating candidate adversarial examples against the (assumed) local surrogates, followed by running the Square Attack on the remaining examples that fail to transfer from, can easily establish a (much) stronger baseline. This idea is inspired by Suya et al. [140], which appears before the ODS-RGF attack [142]. Details of the transfer experiment (on generating local adversarial examples) can be found in Appendix A1. Table II compares ODS-RGF, Square Attack and our proposed Hybrid-Square in terms of the attack success rate and the average number of queries, using the same experimental setup as the original apper [142]. The first two attacks are the proposed and baselinesattacks in Tashiro et al. [142]. We observe that both ODS-RGF and Hybrid-Square improve query efficiency compared to the original Square Attack. However, the Hybrid-Square attack significantly outperforms the proposed ODS-RGF attack, demonstrating the importance of considering simple adaptations of known attacks to new threat models.

At last, stronger baselines may emerge not only when extra information is available but also when attacks utilize auxiliary data, even in the absence of such extra information. As mentioned in Section IV-D, it is worth noting that attacks that operate with auxiliary data can theoretically be applied in settings with varying data sizes. The key distinction lies in the degree of effectiveness these attacks exhibit under different data sizes. Therefore, we recommend that attack methods, which implicitly assume the availability of "sufficient" or "insufficient" auxiliary data, should also use methods from the opposite category as baselines. Furthermore, researchers should conduct ablation studies to examine how the attack performance evolves, compared to the baselines, when transitioning from "insufficient" to "sufficient" auxiliary data.

C. Interaction Among Attacker Knowledge

The most straightforward interaction of attacker knowledge is adversaries can train many pretrained models given enough auxiliary data. Therefore, attacks may treat the existence of sufficient auxiliary data the same as the existence of both the data and the pretrained models (obtained from the data). Further, proper identification of threat models using our taxonomy uncovers connections to other related fields such as model stealing (also known as model extraction) [8] and model inversion [10]. Model stealing adversaries aim to steal a copy of a remotely deployed machine learning model given Oracle prediction access. In contrast, model inversion adversaries seek to infer (parts of) the training distribution of the remote model. These attacks can significantly boost the performance of black-box attacks with interactive access to the target model by providing better surrogate models (via model extraction) and more representative training data (via model inversion). We do not implement these ideas but discuss their potential in detail below.

Model-Extraction Attacks. Simply identifying the target model structure (or family of models) [9, 188] can improve attack success, especially in settings where the auxiliary data highly overlaps with the target model's training data. The extensive literature on attack transferability [33, 42, 58, 64] can thus serve as a "handbook" for adversaries. Further, when we look to utilize model extraction for better transferability of adversarial examples, an adversary's specific goal is to ensure the extracted and victim models have a similar vulnerability space, so that better surrogates can boost black-box attack performance. This is an easier objective that the original model extraction objective of having prediction consistency [189] as it is believed that adversarial examples reside in a low dimensional subspace [9] that is easier to capture than the full input space.

When enough auxiliary data exists, state-of-the-art model extraction attacks can be readily applied [9]. Limited auxiliary data settings are more challenging. Several works on black-box adversarial examples use surrogate training to enhance transferability [171–173] or improve query efficiency [143]. Surrogate training is also common in model extraction [189].

However, these surrogate extraction methods fail for complex image classification tasks. Recent advances in data-free extraction attacks show promise for addressing complex classification tasks and can be further enhanced with pretrained models.

Data-free model-extraction attacks [190–193] rely on a generator to generate queries, which are then labeled by the target model and used to update the generator and the extracted model. These methods work well without any pretrained model—in particular, generators are randomly initialized and optimized with the estimated gradient from the target model [190, 191]. With pretrained models, one may first (pre-)train a generator with auxiliary models (using their actual gradients) and then continue training the generator with estimated gradients from the black-box model. Such a generator is likely better than a randomly initialized one and may enable extraction in fewer queries. The feasibility of pretraining a generator and then fine-tuning for the target model has already been demonstrated when directly generating adversarial examples [87, 183–185]. Additionally, knowledge from the surrogate models may still transfer to the target when the training data of the two models have partial or no overlap [167]. We note that the obtained generator can also be used to augment the adversary's data. When limited quantities of data are available, this increased data can in turn enable other model extraction methods that require more auxiliary data [9].

Model-Inversion Attacks. Model-inversion attacks aim to recover representative and semantically meaningful training data [194] of a given model. However, to generate adversarial examples, the extracted data need not be semantically meaningful [87, 94]. Model inversion can help either directly [87, 94, 95, 183–185] by providing more representative data (which can be further diversified with data augmentations [195]), or indirectly by boosting the performance of model extraction attacks via better query synthesis [195]. In settings with sufficient auxiliary data, state-of-the-art model-inversion attacks [196, 197] can be applied directly to recover more representative data and improve the quality of the auxiliary data. For settings with limited auxiliary data, an adversary may use a query-generator trained during a model-extraction attack (where the generator is a common component in most techniques, as described earlier) to generate more auxiliary data. State-of-the-art black-box inversion attacks [197] can then be utilized in the absence of a surrogate model.

Still motivated by the success of pretraining and fine-tuning generators for adversarial example generation [87, 183–185], we see opportunities for exploiting the presence of pretrained auxiliary models in improving the effectiveness of model inversion attacks against the unknown target, to eventually improve performance of black-box attacks. Particularly, the conditional generative model in Liu et al. [197] can first use labels from auxiliary models, followed by fine-tuning with labels from the target model to improve performance. Similarly, white-box inversion attacks [196] may utilize the auxiliary model for gradient computation and then use predictions from the target model to estimate gradients using black-box gradient estimation [3, 16] techniques for fine-tuning.

Combining Model-Stealing and Model-Inversion. Model stealing and model inversion attacks can be combined dynamically—for instance, by iteratively running model stealing and inversion attacks to boost each other. One thing to note is that these attacks' query requirements can be quite high and unrealistic for resource-constrained adversaries, even though attackers only have to run these two attacks once and then use the results to boost future black-box attacks. For example, even state-of-the-art black-box model inversion attacks require millions of queries (e.g., DiSGUIDE [192] use at least 4M queries for models trained on CIFAR-10 [198]).

VII. Rethinking Baseline Comparisons

Most interactive and non-interactive attacks involve running an optimization loop locally for some number of iterations to find a candidate adversarial example. It is in the adversary's interest to run the attack for as many iterations as possible as long as more iterations improve success rates. The number of iterations is also used as a grounding factor in attack comparisons, running attacks for the same number of iterations for fair comparison [24, 33, 36, 38, 45, 170]. [4] We argue that such measures, while well-meaning, are in fact not "fair" and misaligned with what adversaries care about. Fixing the number of iterations limits some attacks, clipping their potential for the sake of comparison. In most cases, the iteration-wise cost of attacks is low, and an adversary that does not have severe latency requirements would only care about maximizing its success rate. When latency or compute costs matter, an adversary would prefer the attack that yields the highest attack success rate within the given time or resource constraints. As pointed out by Apruzzese et al. [5] through a thorough analysis of real-world adversarial scenarios, attackers prefer cheap and effective methods that can be easily automated, and the relevant cost metric is the total effort spent on the process of completing an attack—a metric that is harder to count than number of iterations, but is more direct.

Another issue with many evaluations is the lack of challenging settings for attack comparisons. Untargeted attacks are much easier than targeted ones; attacking non-robust standard models is easier than attacking adversarially-robust models. Success rates can be very high in easy settings and thus fail to provide useful insights about relative attack performance that can transfer to harder settings. These harder settings are in fact the ones where attacks matter most.

We advocate for evaluating black-box attacks with a realistic consideration of actions that adversaries can take in practice. Specifically, instead of fixating on a specific number of iterations for non-interactive settings (Section VII-A) as the primary metric for comparison, we argue that adversaries should be able to use more iterations when beneficial and the

[4] Ablation studies report the impact of iteration numbers on ASR, but only up to 30 iterations [42]. Studies on higher attack iterations (up to 1,000) do not report iteration-wise results [155].

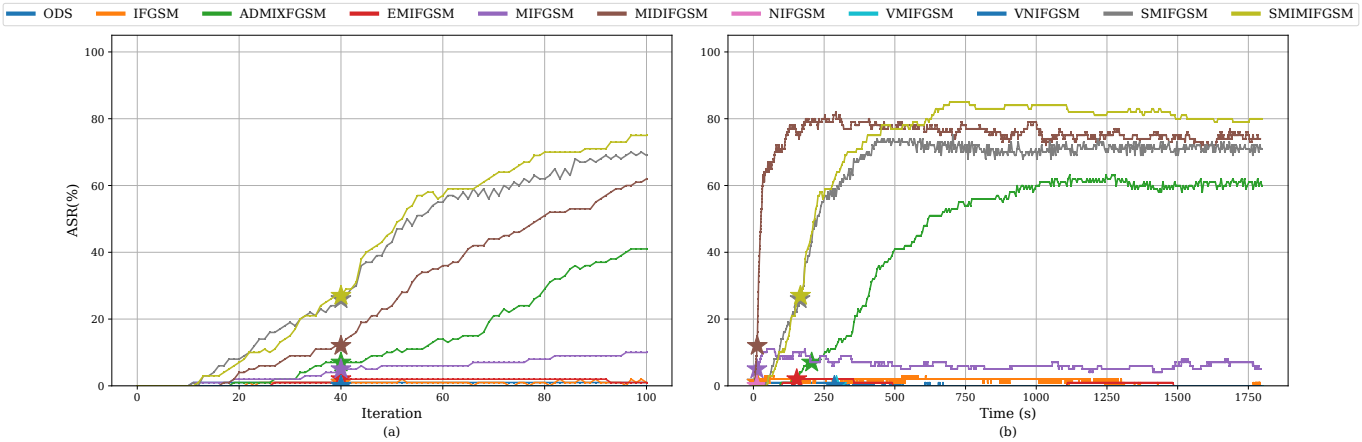

Fig. 2: ASR (y-axis) for various targeted attacks on DenseNet201 models, varying across iterations (a) and time (b). All attacks on the left are run for 100 iterations, while attacks on the right are run for 30 minutes per batch. ASR at each iteration is computed using adversarial examples at that iteration. ASR at 40 iterations are marked with $\star$ for each attack.

only constraints such as total time should be motivated by the evaluation scenario. Our analysis with this new lens uncovers several interesting insights and suggestions for researchers. For non-interactive transfer attacks, we discover how running attacks for more iterations helps attack success (Section VII-A1), and that simply stopping attacks when they succeed on local models can hamper performance (Section VII-A2), and observe much clearer trends in relative attack performance trends when evaluated in hard settings such as targeted attacks (Section VII-A3). In Appendix C, we also show how the ASR of different interactive query-attacks can change when the evaluation metric shifts from the number of queries to the local runtime, and advocate using local runtime as an additional metric on top of the commonly considered query costs for the interactive black-box attacks.

A. Transfer Attacks

Attack success rate (ASR) has been the guiding metric for evaluating different transfer attacks' the effectiveness of different transfer attacks. However, more effective transfer attacks often require complicated computation processes and can lead to local computation costs that are orders of magnitude higher than baselines.

For convenience in comparison, we selected transfer attacks that augment the baseline I-FGSM attack [23] with various gradient and input manipulation techniques, including new combinations (details in Appendix A2)—this leaves us with 20 attacks. Since these attacks are based on iterative local optimizations, we can conveniently measure the impact of different local time constraints on ASR against the target models. Of these 20 attacks, we picked 11 that span a wide range of local runtimes. Note that for each of the following graphs, we re-evaluate the attack at each iteration using the adversarial inputs generated at the end of that iteration, thus giving us multiple attack success rates as iterations progress.

1) Time and iterations: It is conventional to evaluate attacks for a fixed number of iterations: usually 10 for untargeted settings. However, the lack of targeted attack evaluations means there is no such standard for that setting, with MI-FGSM [33] being one of the few attacks that evaluate targeted attacks, using 40 iterations (which is what we set for targeted attacks). However, the number is arbitrary and it is unclear whether attacks have the potential to have improved performance. Prior work [22] also recommends running attacks until convergence, instead of a fixed number of iterations. To test this hypothesis, we run attacks for 100 iterations, instead of the usual 40 for targeted attacks, and analyze ASR trends (Figure 2-a). Most attacks seem to benefit from increased iterations.

Given this potential for improved success beyond 40 iterations, it is important to extend evaluations for valid comparisons. Execution time should be used if resource constraints like runtime exist, especially when the cost-per-iteration varies. Motivated by these factors, we re-run all the attacks but instead of running them for a fixed number of iterations as in prior work, we run them for the same time duration (30 minutes per batch)[5].

Iteration-wise analysis (Figure 2-a) would suggest MIDI-FGSM to be slightly worse off than SMI-FGSM, even when compared under the setting of 100 iterations. However, looking at the same results across time (Figure 2-b), this trend flips once we observe that MIDI-FGSM is nearly 2x as fast and can thus execute double the number of iterations in the same amount of time. Similarly, MIDI-FGSM and Admix-FGSM do not seem very far apart in their performance when looking at the same number of iterations, but time-wise analysis shows how the difference in their performance is much higher.

This analysis based on runtimes paints a clearer picture that is better aligned with what an adversary would desire—maximizing attack success within their available resources (e.g., limit on the total execution time). As an example,

[5]We opt for measuring total runtime over algorithmic measures due to the challenge in standardizing components' runtimes across different hardware configurations, acknowledging both methods have their merits and limitations.

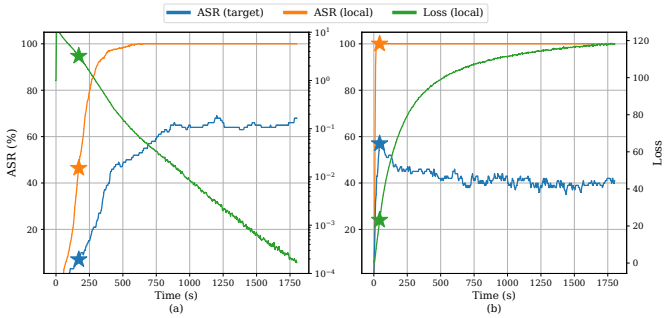

Fig. 3: Attack success rates (ASR) (y-axis, left) for target and local models, along with loss (y-axis, right) while optimizing the objective locally, varying across time (x-axis), for targeted attacks on DenseNet201 (a) and untargeted attacks on adversarially-robust Inception-v3$_{\text{adv}}$ (b), using SMIMI-FGSM [24]. ASR at representative iterations (40 for targeted, 10 for untargeted) are marked with a $\star$ for each of the metrics.

consider an adversary looking at results in the literature to select an attack. Assuming computational constraints are not an issue, the literature would suggest SMI-FGSM being a good candidate instead of candidates like MIDI-FGSM, and the adversary may pick SMI-FGSM to conduct the attack. However, once we realize that these comparisons are based on the number of attack iterations (an arbitrary metric) and instead compare them based on the local runtime, it is clear that MIDI-FGSM can generate better attack results (Figure 2). These are the kind of cases we have in mind while advocating for time-based comparisons. Our motive is not to encourage researchers to add execution-time as an "extra metric", but rather remember that these attacks are designed for adversaries that would only want to maximize success rates given available resources [5], and not care about running the attack for a fixed number of attack iterations.

Recommendation: Run attacks for enough iterations until attack success rates plateau. Execution cost such as the local attack runtime should be used as the equalizing factor when comparing black-box attack performance, not the number of iterations.

2) Knowing When to Stop: As observed in Figure 2, simply running attacks for more iterations often improves attack success rates. For instance, attack success for MIDI-FGSM jumps from $< 20\%$ to nearly 80% when run for sufficient iterations which, interestingly, is still faster than running Admix-FGSM for 40 iterations. Similarly, SMIMI-FGSM jumps from $\sim$75% to $\sim$85%, once the attack runs for longer. However, success rates do not always improve with more iterations. For instance, while MIDI-FGSM in the targeted setting (Figure 2-b) sees an improvement, it fluctuates between 70 and 80%. While running attacks for more iterations helps in most cases, it is not obvious when an adversary should stop their attack to maximize ASR—the adversary cannot know the optimal number of iterations before executing their attack. One possible workaround is keeping track of metrics for the local models (which are used to compute gradients), and possibly running more iterations as long as metrics such as local success rates and loss do not stagnate.

Intuitively, an adversary has no reason to continue local attack optimization once it successfully generates adversarial examples for its local models. The only possible motivation lies in changing the model's prediction probabilities—increasing confidence for targeted attacks, and decreasing confidence for untargeted attacks. Our analysis shows how the rate of finding successful adversarial examples against local models gets to 100% almost immediately, even when attack success rates on the target models are low. An adversary that only inspects local attack success rates would thus stop its optimization prematurely, leading to sub-optimal ASR for the target model.

For the targeted setting (Figure 3-a), we interestingly observe the local loss value to continue dropping (not by much; note that the right y-axis is on log-scale for loss in targeted attack), even though target ASR starts stagnating in the 500-750s range. Looking at such a loss trajectory, it may be tempting to conclude running the attack till the local loss converges, should be a good heuristic for knowing when the target ASR will be highest. However, inspecting the case of an adversarially-robust target model (Figure 3-b) disabuses us of this notion—ASR peaks at around ten iterations, while the local loss keeps increasing and converging until the very end of attack execution. It is not surprising that local loss continues to converge, since this is what the attacks optimize for while computing gradients, and this may not necessarily align well enough with the target model.

The fact that ASR for the target model keeps increasing significantly even after the attack succeeds for local models is intriguing and a challenge unique to black-box attacks. While this goes hand in hand with the suggestion to evaluate attacks for longer iterations, it raises the question of knowing when the attack running locally should be stopped to maximize ASR for the target model.

Recommendation: Do not rely on attack success or loss on local models as a metric to stop optimization. Developing metrics that can help predict optimal target ASR is a direction for future work.

3) Harder settings: Since almost all attacks against standard models with sufficient perturbation budget achieve nearly 100% attack success, there is limited room for improvement. However, attacks in harder settings (Figure 4) can be much less effective (e.g., $< 60\%$ ASR when perturbation budget is halved to 8/255) and can demonstrate different trends in relative performance. For example, against an adversarially trained target model, the least and most performant attacks differ by as much as $\sim$30% in their ASR (similar trends hold for the targeted setting). Although attacks like SMIMI-FGSM seem to perform well across all settings, this is indeed a posterior observation that can only be verified for a new

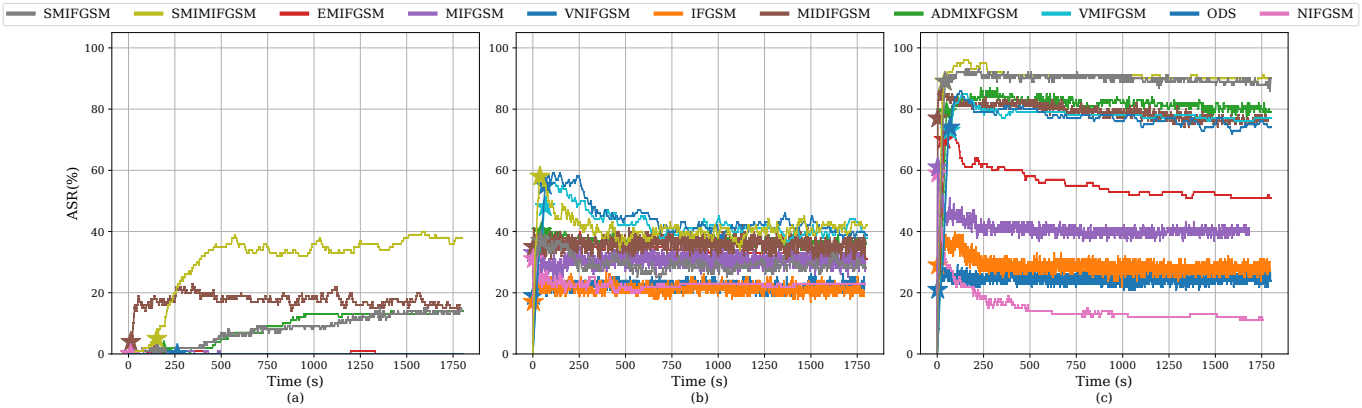

Fig. 4: ASR (y-axis) for various attacks: targeted attacks for Inception-v3 with perturbation budget $16/255$ (ℓ_∞) (a), untargeted attacks for Inception-v3 with reduced perturbation budget $8/255$ (b), and untargeted attacks for adversarially robust model Inc-v3$_{\text{adv}}$ with perturbation budget $16/255$ (c). ASR at each iteration is computed using adversarial examples at that iteration. ASR at representative iterations (40 for targeted, 10 for untargeted) are marked with $\star$ for each attack.

attack when it is evaluated across diverse and hard settings and, in fact, does not hold for other hard settings like IncRes-v2$_{\text{ens}}$ target models (SMIMI-FGSM is out-performed by VMI-FGSM and VNI-FGSM, Figure 5 in the Appendix).

> **Recommendation**: When evaluating and comparing attacks, researchers should include harder attack settings, such as targeted attacks, low perturbation budgets, and adversarially robust target models.

VIII. Discussion

We highlight our key findings, discuss their implications, and make recommendations for future research. We also identify the limitations of this work.

Many Interesting Settings Underexplored. Categorizing attacks from the literature uncovers how several threat models have close to little or no research dedicated to those specific settings (Section VI-A) despite these areas being some of the most relevant to practical attacks—most model APIs return top-k scores (not full confidence vector) and the availability of abundant data from non-overlapping distributions is possible via the Internet, yet both of these settings have hardly any research. We also identify the utility of orthogonal yet useful fields in ML security, such as model extraction and model inversion, and how they can be utilized under certain threat models to boost the performance of black-box attacks (Section VI-C). Future research should focus on developing specific attacks for these unexplored but important and interesting settings.

Careful Evaluation Matters. Even within well-explored threat spaces, researchers often compare proposed attacks with baselines that require different (and often more restrictive) assumptions over the adversary's capabilities, and in settings that are easy enough that all attacks work well. We show how small tweaks to adapt existing methods to utilize the available knowledge fully can strengthen the baselines and outperform the proposed attacks (Section VI-B). Additionally, several attacks focus on the untargeted setting where most attacks already achieve near-perfect ASR, instead of harder settings such as targeted attacks and adversarially robust target models, where attack performance trends can change drastically. We implore researchers to conduct evaluations in settings where differences matter, and to either use state-of-the-art baselines from the same threat space or to adapt baselines to utilize assumed knowledge.

Evaluate Attacks under Well-motivated Constraints. When constraints are imposed on attacks, they should be motivated by realistic adversarial constraints and focus on attack cost. Our experiments demonstrate how several proposed attacks can benefit from more iterations, yet predicting the optimal number of local attack iterations is non-trivial. We thus advocate for a shift in paradigm when reporting attack results for adversarial attacks: using time as the equalizing metric for comparing attacks instead of iterations to infer the attack effectiveness better. We also hope our results motivate future work to use a better selection method for choosing the best candidate examples across iterations.

Limitations. Our analysis of evaluated attacks is focused on image classifiers, which is not a security-critical application. While there are claims that attacks from image classifiers can be adapted to other domains like malware classifiers, there is little evidence that the decisions about which attack to adapt would be based on extensive evaluations in image space.

Acknowledgements

This work was partially funded by awards from the National Science Foundation (NSF) SaTC program (Center for Trustworthy Machine Learning, #1804603), the AI Institute for Agent-based Cyber Threat Intelligence and Operation (ACTION) (#2229876), NSF #2323105, and NSF #2325369.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," in *International Conference on Learning Representations*, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Machine Learning*, 2014.

[3] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box Adversarial Attacks with Limited Queries and Information," in *International Conference on Machine Learning*. PMLR, 2018.

[4] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms," in *European Conference on Computer Vision*, 2018.

[5] G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2023.

[6] S. Bhambri, S. Muku, A. Tulasi, and A. B. Buduru, "A Survey of Black-Box Adversarial Attacks on Computer Vision Models," *arXiv:1912.01667*, 2019.

[7] K. Mahmood, R. Mahmood, E. Rathbun, and M. Van Dijk, "Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks," *IEEE Access*, 2021.

[8] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs," in *USENIX Security Symposium*, 2016.

[9] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High Accuracy and High Fidelity Extraction of Neural Networks," in *USENIX Security Symposium*, 2020.

[10] M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures," in *The ACM Conference on Computer and Communications Security*, 2015.

[11] T. Wang, Y. Zhang, and R. Jia, "Improving Robustness to Model Inversion Attacks via Mutual Information Regularization," in *The AAAI Conference on Artificial Intelligence*, 2021.

[12] A. Mehra, B. Kailkhura, P.-Y. Chen, and J. Hamm, "How Robust are Randomized Smoothing Based Defenses to Data Poisoning?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[13] E. Radiya-Dixit and F. Tramèr, "Data Poisoning Won't Save You From Facial Recognition," in *International Conference on Learning Representations*, 2022.

[14] D. I. Diochnos, S. Mahloujifar, and M. Mahmoody, "Lower bounds for adversarially robust PAC learning under evasion and hybrid attacks," in *IEEE International Conference on Machine Learning and Applications*, 2020.

[15] D. Diochnos, S. Mahloujifar, and M. Mahmoody, "Adversarial risk and robustness: General definitions and implications for the uniform distribution," *Advances in Neural Information Processing Systems*, 2018.

[16] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models," in *10th ACM workshop on artificial intelligence and security*, 2017.

[17] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *IEEE Symposium on Security and Privacy*, 2017.

[18] Z. Zhao, H. Zhang, R. Li, R. Sicre, L. Amsaleg, and M. Backes, "Towards Good Practices in Evaluating Transfer Adversarial Attacks," *arXiv:2211.09565*, 2022.

[19] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems," in *IEEE Symposium on Security and Privacy*, 2021.

[20] L. Li, X. Qi, T. Xie, and B. Li, "SoK: Certified Robustness for Deep Neural Networks," *IEEE Symposium on Security and Privacy*, 2020.

[21] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and Privacy in Machine Learning," in *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

[22] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On Evaluating Adversarial Robustness," *arXiv preprint arXiv:1902.06705*, 2019.

[23] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," in *International Conference on Machine Learning*, 2016.

[24] G. Wang, H. Yan, and X. Wei, "Enhancing Transferability of Adversarial Examples with Spatial Momentum," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2022.

[25] L. Wu, Z. Zhu, C. Tai *et al.*, "Understanding and Enhancing the Transferability of Adversarial Examples," *arXiv:1802.09707*, 2018.

[26] L. Gao, Q. Zhang, J. Song, X. Liu, and H. T. Shen, "Patch-wise Attack for Fooling Deep Neural Network," in *European Conference on Computer Vision*. Springer, 2020.

[27] L. Gao, Q. Zhang, J. Song, and H. T. Shen, "Patch-wise++ Perturbation for Adversarial Targeted Attacks," *CoRR*, 2020.

[28] Y. Li, S. Bai, C. Xie, Z. Liao, X. Shen, and A. Yuille, "Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses," in *European Conference on Computer Vision*. Springer, 2020.

[29] L. Gao, Q. Zhang, X. Zhu, J. Song, and H. T. Shen, "Staircase Sign Method for Boosting Adversarial Attacks," *CoRR*, 2021.

[30] H. Tan, Z. Gu, L. Wang, H. Zhang, B. B. Gupta, and Z. Tian, "Improving Adversarial Transferability by Temporal and Spatial Momentum in Urban Speaker Recognition Systems," *Computers and Electrical Engineering*, 2022.

[31] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature Importance-aware Transferable Adversarial Attacks," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[32] S. Fang, J. Li, X. Lin, and R. Ji, "Learning to Learn Transferable Attack," in *AAAI Conference on Artificial Intelligence*, 2022.

[33] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting Adversarial Attacks with Momentum," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[34] 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.

[35] J. Zou, Y. Duan, B. Li, W. Zhang, Y. Pan, and Z. Pan, "Making Adversarial Examples More Transferable and Indistinguishable," in *AAAI Conference on Artificial Intelligence*, 2022.

[36] X. Wang, J. Lin, H. Hu, J. Wang, and K. He, "Boosting Adversarial Transferability through Enhanced Momentum," in *British Machine Vision Conference*, 2021.

[37] D. Jang, S. Son, and D.-S. Kim, "Strengthening the Transferability of Adversarial Examples Using Advanced Looking Ahead and Self-CutMix," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

[38] X. Wang and K. He, "Enhancing the Transferability of Adversarial Attacks through Variance Tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[39] Q. Lu, S. Wei, H. Chu, and Y. Zhao, "Towards Transferable 3D Adversarial Attack," in *ACM Multimedia Asia*. Association for Computing Machinery, 2021.

[40] Z. He, Y. Duan, W. Zhang, J. Zou, Z. He, Y. Wang, and Z. Pan, "Boosting Adversarial Attacks with Transformed Gradient," *Computers & Security*, 2022.

[41] K. Liang, J. Y. Zhang, B. Wang, Z. Yang, S. Koyejo, and B. Li, "Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability ," in *International Conference on Machine Learning*. PMLR, 2021.

[42] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[43] J. Zou, Z. Pan, J. Qiu, X. Liu, T. Rui, and W. Li, "Improving the Transferability of Adversarial Examples with Resized-Diverse-Inputs, Diversity-Ensemble and Region Fitting," in *European Conference on Computer Vision*. Springer, 2020.

[44] B. Yang, H. Zhang, Z. Li, Y. Zhang, K. Xu, and J. Wang, "Adversarial example generation with AdaBelief Optimizer and Crop Invariance," *Applied Intelligence*, 2022.

[45] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[46] 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*. Springer, 2022.

[47] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, "Improving the transferability of targeted adversarial examples through object-based diverse input," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[48] X. Wang, Z. Zhang, and J. Zhang, "Structure Invariant Transformation for better Adversarial Transferability," in *IEEE/CVF International Conference on Computer Vision*, 2023.

[49] W. Wu, Y. Su, M. R. Lyu, and I. King, "Improving the Transferability of Adversarial Samples with Adversarial Transformations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[50] Z. Yuan, J. Zhang, and S. Shan, "Adaptive Image Transformations for Transfer-based Adversarial Attack," *European Conference on Computer Vision*, 2021.

[51] J. Zhang, J.-t. Huang, W. Wang, Y. Li, W. Wu, X. Wang, Y. Su, and M. R. Lyu, "Improving the transferability of adversarial samples by path-augmented method," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[52] H. Yuan, Q. Chu, F. Zhu, R. Zhao, B. Liu, and N.-H. Yu, "AutoMA: Towards Automatic Model Augmentation for Transferable Adversarial Attacks," *IEEE Transactions on Multimedia*, 2021.

[53] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, "Towards Transferable Targeted Attack," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[54] Z. Zhao, Z. Liu, and M. Larson, "On Success and Simplicity: A Second Look at Transferable Targeted Attacks," in *Advances in Neural Information Processing Systems*, 2021.

[55] 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.

[56] X. Wang, J. Ren, S. Lin, X. Zhu, Y. Wang, and Q. Zhang, "A Unified Approach to Interpreting and Boosting Adversarial Transferability," *International Conference on Learning Representations*, 2020.

[57] A. Ganeshan, V. BS, and R. V. Babu, "FDA: Feature Disruptive Attack," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[58] N. Inkawhich, K. J. Liang, L. Carin, and Y. Chen, "Transferable Perturbations of Deep Feature Distributions," in *International Conference on Machine Learning*, 2020.

[59] N. Inkawhich, K. Liang, B. Wang, M. Inkawhich, L. Carin, and Y. Chen, "Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability," *Advances in Neural Information Processing Systems*, 2020.

[60] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing Adversarial Example Transferability with an Intermediate Level Attack," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[61] Q. Li, Y. Guo, and H. Chen, "Yet Another Intermediate-Level Attack," in *European Conference on Computer Vision*. Springer, 2020.

[62] Z. Liu, Z. Zhao, and M. Larson, "Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval," in *International Conference on Multimedia Retrieval*, 2019.

[63] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable Adversarial Perturbations," in *European Conference on Computer Vision*, 2018.

[64] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature Space Perturbations Yield More Transferable Adversarial Examples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[65] W. Wu, Y. Su, X. Chen, S. Zhao, I. King, M. R. Lyu, and Y.-W. Tai, "Boosting the Transferability of Adversarial Samples via Attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[66] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[67] R. Wang, Y. Guo, R. Yang, and Y. Wang, "Exploring Transferable and Robust Adversarial Perturbation Generation from the Perspective of Network Hierarchy," *arXiv preprint arXiv:2108.07033*, 2021.

[68] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of Neural Network Representations Revisited," in *International Conference on Machine Learning*. PMLR, 2019.

[69] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014.

[70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *IEEE/CVF International Conference on Computer Vision*, 2017.

[71] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *International Conference on Machine Learning*. PMLR, 2017.

[72] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into Transferable Adversarial Examples and Black-box Attacks," *International Conference on Machine Learning*, 2017.

[73] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning Transferable Adversarial Examples via Ghost Networks," in *AAAI Conference on Artificial Intelligence*, 2020.

[74] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets," in *International Conference on Learning Representations*, 2020.

[75] Y. Duan, J. Zou, X. Zhou, W. Zhang, Z. He, D. Zhan, J. Zhang, and Z. Pan, "Adversarial Attack via Dual-Stage Network Erosion," *Computers & Security*, 2022.

[76] Z. Xu, Z. Gu, J. Zhang, S. Cui, C. Meng, and W. Wang, "Back-propagation path search on adversarial transferability," in *IEEE/CVF International Conference on Computer Vision*, 2023.

[77] Y. Zhu, J. Sun, and Z. Li, "Rethinking Adversarial Transferability from a Data Distribution Perspective," in *International Conference on Machine Learning*, 2021.

[78] C. Zhang, P. Benz, G. Cho, A. Karjauv, S. Ham, C.-H. Youn, and I. S. Kweon, "Backpropagating smoothly improves transferability of adversarial examples," in *CVPR 2021 Workshop Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges (AML-CV)*, vol. 2, 2021.

[79] Y. Guo, Q. Li, and H. Chen, "Backpropagating Linearly Improves Transferability of Adversarial Examples," in *Advances in Neural Information Processing Systems*, 2020.

[80] X. Wang, K. Tong, and K. He, "Rethinking the backward propagation for adversarial transferability," in *Advances in Neural Information Processing Systems*, 2023.

[81] H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry, "Do Adversarially Robust ImageNet Models Transfer Better?" *Advances in Neural Information Processing Systems*, 2020.

[82] F. Utrera, E. Kravitz, N. B. Erichson, R. Khanna, and M. W. Mahoney, "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification," in *International Conference on Learning Representations*, 2021.

[83] Z. Deng, L. Zhang, K. Vodrahalli, K. Kawaguchi, and J. Y. Zou, "Adversarial Training Helps Transfer Learning via Better Representations," in *Advances in Neural Information Processing Systems*, 2021.

[84] F. Zhou, H. Ling, Y. Shi, J. Chen, Z. Li, and Q. Wang, "Improving Transferability of Adversarial Examples on Face Recognition with Beneficial Perturbation Feature Augmentation," *arXiv:2210.16117*, 2022.

[85] J. Springer, M. Mitchell, and G. Kenyon, "A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks," in *Advances in Neural Information Processing Systems*, 2021.

[86] C. Zhang, G. Cho, P. Benz, K. Zhang, C. Zhang, C.-H. Youn, and I. S. Kweon, "Early Stop And Adversarial Training Yield Better surrogate Model: Very Non-Robust Features Harm Adversarial Transferability," *OpenReview*, 2021.

[87] X. Yang, Y. Dong, T. Pang, H. Su, and J. Zhu, "Boosting Transferability of Targeted Adversarial Examples via Hierarchical Generative Networks," in *European Conference on Computer Vision*. Springer, 2022.

[88] Z. Yuan, J. Zhang, Y. Jia, C. Tan, T. Xue, and S. Shan, "Meta Gradient Adversarial Attack," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[89] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative Adversarial Perturbations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[90] M. M. Naseer, S. H. Khan, M. H. Khan, F. Shahbaz Khan, and F. Porikli, "Cross-Domain Transferability of Adversarial Perturbations," *Advances in Neural Information Processing Systems*, 2019.

[91] K. kanth Nakka and M. Salzmann, "Learning Transferable Adversarial Perturbations," in *Advances in Neural Information Processing Systems*, 2021.

[92] Q. Zhang, X. Li, Y. Chen, J. Song, L. Gao, Y. He, and H. Xue, "Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains," in *International Conference on Learning Representations*, 2022.

[93] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "On Generating Transferable Targeted Perturbations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[94] J. Zhang, L. Li, H. Li, X. Zhang, S. Yang, and B. Li, "Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation," in *International Conference on Machine Learning*. PMLR, 2021.

[95] H. Li, L. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "Nonlinear Projection Based Gradient Estimation for Query Efficient Blackbox Attacks," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

[96] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square Attack: a query-efficient black-box adversarial attack via random search," in *European Conference on Computer Vision*, 2020.

[97] J. Chen and Q. Gu, "RayS: A Ray Searching Method for Hard-label Adversarial Attack," in *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[98] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach," in *International Conference on Machine Learning*, 2018.

[99] M. Cheng, S. Singh, P. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh, "Sign-OPT: A Query-Efficient Hard-label Adversarial Attack," in *International Conference on Machine Learning*, 2019.

[100] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A Query-Efficient Decision-Based Attack," in *IEEE Symposium on Security and Privacy*, 2020.

[101] Y. Du, M. Fang, J. Yi, J. Cheng, and D. Tao, "Towards Query Efficient Black-box Attacks: An Input-free Perspective," in *11th ACM Workshop on Artificial Intelligence and Security*, 2018.

[102] S. Liu, P.-Y. Chen, X. Chen, and M. Hong, "signSGD via Zeroth-Order Oracle," in *International Conference on Machine Learning*, 2018.

[103] J. Uesato, B. O'donoghue, P. Kohli, and A. Oord, "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," in *International Conference on Machine Learning*. PMLR, 2018.

[104] J. Li, R. Ji, H. Liu, J. Liu, B. Zhong, C. Deng, and Q. Tian, "Projection & Probability-Driven Black-Box Attack," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[105] A. Ilyas, L. Engstrom, and A. Madry, "Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors," in *International Conference on Machine Learning*, 2018.

[106] Y. Shi, S. Wang, and Y. Han, "Curls & Whey: Boosting Black-Box Adversarial Attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[107] P. Zhao, P.-Y. Chen, S. Wang, and X. Lin, "Towards Query-Efficient Black-Box Adversary with Zeroth-Order Natural Gradient Descent," in *AAAI Conference on Artificial Intelligence*, 2020.

[108] P. Zhao, S. Liu, P.-Y. Chen, N. Hoang, K. Xu, B. Kailkhura, and X. Lin, "On the Design of Black-box Adversarial Examples by Leveraging Gradient-free Optimization and Operator Splitting Method," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[109] J. Chen, D. Zhou, J. Yi, and Q. Gu, "A Frank-Wolfe Framework for Efficient and Effective Adversarial Attacks," in *AAAI conference on artificial intelligence*, 2020.

[110] A. Al-Dujaili and U.-M. O'Reilly, "Sign Bits Are All You Need for Black-Box Attacks," in *International Conference on Machine Learning*, 2019.

[111] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "NATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks," in *International Conference on Machine Learning*. PMLR, 2019.

[112] C.-J. Simon-Gabriel, N. A. Sheikh, and A. Krause, "PopSkipJump: Decision-Based Attack for Probabilistic Classifiers," in *38th International Conference on Machine Learning*. PMLR, 2021.

[113] D. Wang, J. Lin, and Y.-G. Wang, "Query-Efficient Adversarial Attack Based on Latin Hypercube Sampling," in *IEEE International Conference on Image Processing*. IEEE, 2022.

[114] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai, "GeoDA: a geometric framework for black-box adversarial attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[115] Y. Liu, S.-M. Moosavi-Dezfooli, and P. Frossard, "A geometry-inspired decision-based attack," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[116] W. Zhao and Z. Zeng, "Improved black-box attack based on query and perturbation distribution," in *13th International Conference on Advanced Computational Intelligence*. IEEE, 2021.

[117] J. Chen, M. Su, S. Shen, H. Xiong, and H. Zheng, "POBA-GA: Perturbation Optimized Black-Box Adversarial Attacks via Genetic Algorithm," *Computers & Security*, 2019.

[118] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "GenAttack: Practical Black-box Attacks with Gradient-Free Optimization," in *The Genetic and Evolutionary Computation Conference*, 2019.

[119] L. Meunier, J. Atif, and O. Teytaud, "Yet another but more efficient black-box adversarial attack: tiling and evolution strategies," *arXiv:1910.02244*, 2019.

[120] B. Ru, A. Cobb, A. Blaas, and Y. Gal, "Bayesopt Adversarial Attack," in *International Conference on Machine Learning*, 2019.

[121] F. Suya, Y. Tian, D. Evans, and P. Papotti, "Query-limited Black-box Attacks to Classifiers," in *NIPS Workshop on Machine Learning and Computer Security Workshop*, 2017.

[122] S. Moon, G. An, and H. O. Song, "Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization," in *International Conference on Machine Learning*. PMLR, 2019.

[123] N. Narodytska and S. P. Kasiviswanathan, "Simple Black-Box Adversarial Attacks on Deep Neural Networks," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[124] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple Black-box Adversarial Attacks," in *International Conference on Machine Learning*. PMLR, 2019.

[125] H. Tran, D. Lu, and G. Zhang, "Exploiting the Local Parabolic Landscapes of Adversarial Losses to Accelerate Black-Box Adversarial Attack," in *European Conference on Computer Vision*. Springer, 2022.

[126] F. Croce and M. Hein, "Sparse and Imperceivable Adversarial Attacks," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[127] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein, "Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks," in *AAAI Conference on Artificial Intelligence*, 2022.

[128] N. Shiva Kasiviswanathan *et al.*, "imple Black-Box Adversarial Attacks on Deep Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[129] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," in *International Conference on Machine Learning*, 2017.

[130] T. Brunner, F. Diehl, M. T. Le, and A. Knoll, "Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks," in *IEEE/CVF International Conference on Computer Vision*, 2019.

[131] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient Decision-based Black-box Adversarial Attacks on Face Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[132] Y. Shi, Y. Han, and Q. Tian, "Polishing Decision-Based Adversarial Noise With a Customized Sampling," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[133] J. Li, R. Ji, P. Chen, B. Zhang, X. Hong, R. Zhang, S. Li, J. Li, F. Huang, and Y. Wu, "Aha! Adaptive History-driven Attack for Decision-based Black-box Models," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[134] X. Sun, G. Cheng, L. Pei, and J. Han, "Query-efficient decision-based attack via sampling distribution reshaping," *Pattern Recognition*, 2022.

[135] T. Maho, T. Furon, and E. Le Merrer, "SurFree: a fast surrogate-free black-box attack," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[136] W. Chen, Z. Zhang, X. Hu, and B. Wu, "Boosting Decision-Based Black-Box Adversarial Attacks with Random Sign Flip," in *European Conference on Computer Vision*. Springer, 2020.

[137] K. A. Midtlid, J. Åsheim, and J. Li, "Magnitude Adversarial Spectrum Search-based Black-box Attack against Image Classification," in *15th ACM Workshop on Artificial Intelligence and Security*, 2022.

[138] V. Q. Vo, E. Abbasnejad, and D. C. Ranasinghe, "Query Efficient Decision Based Sparse Attacks Against Black-Box Deep Learning Models," in *International Conference on Learning Representations*, 2022.

[139] X. Wang, Z. Zhang, K. Tong, D. Gong, K. He, Z. Li, and W. Liu, "Triangle Attack: A Query-Efficient Decision-Based Adversarial Attack," in *European Conference on Computer Vision*. Springer, 2022.

[140] F. Suya, J. Chi, D. Evans, and Y. Tian, "Hybrid Batch Attacks: Finding Black-box Adversarial Examples with Limited Queries," in *USENIX Security Symposium*, 2020.

[141] Z. Huang and T. Zhang, "Black-Box Adversarial Attack with Transferable Model-based Embedding," in *International Conference on Learning Representations*, 2020.

[142] Y. Tashiro, Y. Song, and S. Ermon, "Diversity can be Transferred: Output Diversification for White- and Black-box Attacks," in *Advances in Neural Information Processing Systems*, 2020.

[143] J. Yang, Y. Jiang, X. Huang, B. Ni, and C. Zhao, "Learning Black-Box Attackers with Transferable Priors and Query Feedback," in *Advances in Neural Information Processing Systems*, 2020.

[144] N. A. Lord, R. Mueller, and L. Bertinetto, "Attacking deep networks with surrogate-based adversarial black-box methods is easy," in *International Conference on Learning Representations*, 2022.

[145] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving Black-box Adversarial Attacks with a Transfer-based Prior," in *Advances in Neural Information Processing Systems*, 2019.

[146] Y. Guo, Z. Yan, and C. Zhang, "Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks," in *Advances in Neural Information Processing Systems*, 2019.

[147] C. Ma, S. Cheng, L. Chen, J. Zhu, and J. Yong, "Switching Transferable Gradient Directions for Query-Efficient Black-Box Adversarial Attacks," *arXiv:2009.07191*, 2020.

[148] Z. Chen, J. Ding, F. Wu, C. Zhang, Y. Sun, J. Sun, S. Liu, and Y. Ji, "An Optimized Black-Box Adversarial Simulator Attack Based on Meta-Learning," *Entropy*, 2022.

[149] S. Chen, Z. Huang, Q. Tao, and X. Huang, "QueryNet: Attack by Multi-Identity Surrogates," *arXiv:2105.15010*, 2021.

[150] Z. Cai, C. Song, S. Krishnamurthy, A. Roy-Chowdhury, and M. S. Asif, "Blackbox Attacks via Surrogate Ensemble Search," in *Advances in Neural Information Processing Systems*, 2022.

[151] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.

[152] G. Severi, W. Pearce, and A. Oprea, "Bad citrus: Reducing adversarial costs with model distances," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2022.

[153] J. Hou, H. Liu, Y. Liu, Y. Wang, P.-J. Wan, and X.-Y. Li, "Model Protection: Real-time Privacy-preserving Inference Service for Model Privacy at the Edge," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[154] Z. Gu, H. Huang, J. Zhang, D. Su, H. Jamjoom, A. Lamba, D. Pendarakis, and I. Molloy, "YerbaBuena: Securing Deep Learning Inference Data via Enclave-based Ternary Model Partitioning," *arXiv preprint arXiv:1807.00969*, 2018.

[155] L. E. Richards, A. Nguyen, R. Capps, S. Forsyth, C. Matuszek, and E. Raff, "Adversarial Transfer Attacks With Unknown Data and Class Overlap," in *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, 2021.

[156] Q. Zhang, C. Zhang, C. Li, J. Song, L. Gao, and H. T. Shen, "Practical No-box Adversarial Attacks with Training-free Hybrid Image Transformation," *arXiv:2203.04607*, 2022.

[157] Z. Huan, Y. Wang, X. Zhang, L. Shang, C. Fu, and J. Zhou, "Data-Free Adversarial Perturbations for Practical Black-Box Attack," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2020.

[158] N. Inkawhich, K. J. Liang, J. Zhang, H. Yang, H. Li, and Y. Chen, "Can Targeted Adversarial Examples Transfer When the Source and Target Models Have No Label Space Overlap?" in *IEEE/CVF International Conference on Computer Vision*, 2021.

[159] Y. Lu, Y. Jia, J. Wang, B. Li, W. Chai, L. Carin, and S. Velipasalar, "Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[160] K. R. Mopuri, A. Ganeshan, and R. V. Babu, "Generalizable Data-free Objective for Crafting Universal Adversarial Perturbations," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[161] K. R. Mopuri, U. Garg, and R. V. Babu, "Fast Feature Fool: A data independent approach to universal adversarial perturbations," *arXiv:1707.05572*, 2017.

[162] M. Naseer, S. H. Khan, S. Rahman, and F. Porikli, "Task-generalizable Adversarial Attack based on Perceptual Metric," *arXiv preprint arXiv:1811.09020*, 2018.

[163] Y. Qin, Y. Xiong, J. Yi, and C.-J. Hsieh, "Adversarial Attack across Datasets," *arXiv:2110.07718*, 2021.

[164] J. Wu, M. Zhou, S. Liu, Y. Liu, and C. Zhu, "Decision-based Universal Adversarial Attack," *arXiv:2009.07024*, 2020.

[165] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, "Data-Free Adversarial Perturbations for Practical Black-Box Attack," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[166] S. N. Shukla, A. K. Sahu, D. Willmott, and Z. Kolter, "Simple and Efficient Hard Label Black-box Adversarial Attacks in Low Query Budget Regimes," in *27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

[167] M. Yatsura, J. Metzen, and M. Hein, "Meta-Learning the Search Distribution of Black-Box Random Search Based Adversarial Attacks," *Advances in Neural Information Processing Systems*, 2021.

[168] Q. Li, Y. Guo, and H. Chen, "Practical No-box Adversarial Attacks against DNNs," in *Advances in Neural Information Processing Systems*, 2020.

[169] C. Sun, Y. Zhang, W. Chaoqun, Q. Wang, Y. Li, T. Liu, B. Han, and X. Tian, "Towards Lightweight Black-Box Attacks against Deep Neural Networks," in *Advances in Neural Information Processing Systems*, 2022.

[170] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the Transferability of Adversarial Attacks," in *IEEE/CVF International Conference on Computer Vision*, 2021.

[171] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples," *arXiv:1605.07277*, 2016.

[172] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks against Machine Learning," in *ACM on Asia conference on computer and communications security*, 2017.

[173] L. Pengcheng, J. Yi, and L. Zhang, "Query-Efficient Black-box Attack by Active Learning," in *IEEE International Conference on Data Mining*, 2018.

[174] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "DaST: Data-free Substitute Training for Adversarial Attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[175] L. Wang, H. Zhang, J. Yi, C.-J. Hsieh, and Y. Jiang, "Spanning attack: reinforce black-box attacks with unlabeled data," *Machine Learning*, 2020.

[176] Y. Feng, B. Wu, Y. Fan, L. Liu, Z. Li, and S.-T. Xia, "Boosting Black-Box Attack with Partially Transferred Conditional Adversarial Distribution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[177] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng, "AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks," in *AAAI Conference on Artificial Intelligence*, 2019.

[178] H. Mohaghegh Dolatabadi, S. Erfani, and C. Leckie, "AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows," *Advances in Neural Information Processing Systems*, 2020.

[179] Y. Bai, Y. Zeng, Y. Jiang, Y. Wang, S.-T. Xia, and W. Guo, "Improving Query Efficiency of Black-box Adversarial Attack," in *European Conference on Computer Vision*, 2020.

[180] S. Baluja and I. Fischer, "Adversarial Transformation Networks: Learning to Generate Adversarial Examples," *arXiv:1703.09387*, 2017.

[181] J. Bose, G. Gidel, H. Berard, A. Cianflone, P. Vincent, S. Lacoste-Julien, and W. Hamilton, "Adversarial Example Game," in *Advances in Neural Information Processing Systems*, 2020.

[182] A. S. Hashemi, A. Bär, S. Mozaffari, and T. Fingscheidt, "Transferable Universal Adversarial Perturbations Using Generative Models," *arXiv:2010.14919*, 2020.

[183] J. Du, H. Zhang, J. T. Zhou, Y. Yang, and J. Feng, "Query-efficient Meta Attack to Deep Neural Networks," in *International Conference on Machine Learning*, 2019.

[184] C. Ma, L. Chen, and J.-H. Yong, "Simulating Unknown Target Models for Query-Efficient Black-box Attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[185] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating Adversarial Examples with Adversarial Networks," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.
[186] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Adversarial Attack and Defense for Medical Image Analysis: Methods and Applications," *arXiv preprint arXiv:2303.14133*, 2023.
[187] S. Hu, L. Ke, X. Wang, and S. Lyu, "TkML-AP: Adversarial Attacks to Top-k Multi-Label Learning," in *ICCV*, 2021.
[188] S. J. Oh, B. Schiele, and M. Fritz, "Towards Reverse-Engineering Black-Box Neural Networks," *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019.
[189] D. Oliynyk, R. Mayer, and A. Rauber, "I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences," *ACM Computing Surveys*, 2023.
[190] S. Kariyappa, A. Prakash, and M. K. Qureshi, "MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[191] J.-B. Truong, P. Maini, R. J. Walls, and N. Papernot, "Data-Free Model Extraction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
[192] J. Rosenthal, E. Enouen, H. V. Pham, and L. Tan, "DisGUIDE: Disagreement-Guided Data-Free Model Extraction," in *The AAAI Conference on Artificial Intelligence*, 2023.
[193] S. Sanyal, S. Addepalli, and R. V. Babu, "Towards Data-Free Model Stealing in a Hard Label Setting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
[194] S. Chen, M. Kahla, R. Jia, and G.-J. Qi, "Knowledge-Enriched Distributional Model Inversion Attacks," in *IEEE/CVF International Conference on Computer Vision*, 2021.
[195] X. Gong, Y. Chen, W. Yang, G. Mei, and Q. Wang, "InverseNet: Augmenting Model Extraction Attacks with Training Data Inversion." in *IJCAI*, 2021.
[196] X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang, "Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network," in *The AAAI Conference on Artificial Intelligence*, 2023.
[197] R. Liu, "Unstoppable Attack: Label-Only Model Inversion via Conditional Diffusion Model," in *The ACM Conference on Computer and Communications Security*, 2023.
[198] A. Krizhevsky *et al.*, "Learning Multiple Layers of Features from Tiny Images," 2009.
[199] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.
[200] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
[201] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.
[202] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[203] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of Torch," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010.
[204] A. Kurakin, I. J. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, J. Wang, Z. Zhang, Z. Ren, A. L. Yuille, S. Huang, Y. Zhao, Y. Zhao, Z. Han, J. Long, Y. Berdibekov, T. Akiba, S. Tokui, and M. Abe, "Adversarial Attacks and Defences Competition," in *The NIPS '17 Competition: Building Intelligent Systems*. Springer, 2018.
[205] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.
[206] S. Chen, N. Carlini, and D. Wagner, "Stateful Detection of Black-Box Adversarial Attacks," in *1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2020.

APPENDIX

A. Experiments

1) Implementation Details: Below, we provide details around our experimental setup and evaluations.

Models. For the normal setting, we consider DenseNet201 [199], Inception-v3 [200], Resnet101 [201], and VGG19 [202] as the target models. All of these normally-trained models were used from the Torchvision [203] library. Additionally, we also consider two robust target models: Inc-v3$_{\text{adv}}$ [204] and IncRes-v2$_{\text{ens}}$ [204]. For all of our attacks, our local (surrogate) models consist of an ensemble of: DenseNet-121 [199], Inception-v4 [200], ResNet-50 [201], and VGG-16 [202], which do not overlap with the target models.

Data. We randomly sampled 100 images from the ImageNet [205] validation set. To avoid confusion between the two definitions of untargeted attacks (flipping the model's prediction, or making the prediction mismatch the ground-truth), we picked these 100 examples such that all target models have 100% classification accuracy on them.

Attacks. Unless explicitly specified otherwise, all attacks are generated under $16/255$ ℓ_∞ perturbation budget, and hyper-parameters are adopted from the original papers for each of the attacks. The typical setting of step size α is set as ϵ/T, where T is number of iterations, and is set as 40 for targeted attacks, and 10 for untargeted attacks.

Code/Experiments. We used a batch-size of 5 across all transfer attack experiments to make sure that all attacks (given their varying GPU memory requirements) can fit on the GPU for any given batch. All of our experiments we performed on a 2 CPU, 8-core (2 threads/core) CPU, with 64GB RAM and an Nvidia GTX1080Ti server with 11GB memory, running on Ubuntu Server 22.04. All of our attacks were implemented using PyTorch 1.12.1, running on Python 3.7.13. We exclusively run one experiment at a time on the machine while, although time consuming, helps calculate accurate runtime estimates of attacks without potential fluctuations or slowdowns because of other jobs possibly running on the same machine.

2) Attacks Evaluated in This Paper: Below, we provide brief details about the attacks used for evaluations in Section VII.

Non-interactive Transfer Attacks. Fast Gradient Sign Method (FGSM) [2] generates input perturbation by adding noise in the direction of the sign of gradient of the loss with respect to the input image. I-FGSM (Iterative FGSM) [23] is an iterative version of FGSM that applies the FGSM with smaller step size for multiple iterations and strengthens the effectiveness. I-FGSM also becomes the building block of stronger attacks incorporate additional information. For input augmentation methods, Admix-FGSM [170] augments the input of I-FGSM by adding a small patch from other images. ODS-FGSM [142] introduces a sampling strategy for the generated adversarial examples to prioritize diversity in the target model's outputs and improves transferability. The rest of the described attacks enhance the performance of I-FGSM with gradient stabilization. MI-FGSM (Momentum Iterative FGSM) [33] enhances I-FGSM by incorporating momentum in gradient calculation while NI-FGSM [34] uses Nesterov accelerated gradient for I-FGSM to effectively look ahead and improve performance. VMI-FGSM [38] and VNI-FGSM [38] respectively further stabilize the MI-FGSM and NI-FGSM method by incorporating variance of previous gradients. SMI-FGSM [24] considers the (spatial) context gradient information from different regions of the image for stabilization while SMIMI-FGSM from the same paper further augments it by adding temporal momentum. EMI-FGSM [36] considers the average gradient of data points sampled in the gradient direction from previous iterations.

Query-based Interactive Attacks. Bayesian optimization with perturbation sampling from a low dimensional space is leveraged to improve the query efficiency of black-box attacks in the low-query regime, for both the full-score (complete confidence vector) [120] and the hard-label settings [166]. The bayesian optimization based attacks can be efficient in the low-query regime as it judiciously chooses the next sample to query based on a proper modeling of the adversarial space distributed around the victim image. However, this attack cannot scale to larger number of queries because the associated Gaussian process will need to maintain a very large kernel matrix, and make the attack extremely slow to optimize and consume huge memory at high number of queries. Some efficient random search based strategies are also proposed for the full-score [96] and hard-label [97] attacks. Although these attacks are not particularly designed for the low-query regime, they are very efficient to run locally and also shows competitive attack success rate in different query regimes (especially for very high number of queries).

*3) top-*k *Adaptation Details:* For untargeted attacks, full-score attacks can be applied directly to the top-*k* setting—most of these attacks only require the prediction score of the ground-truth class, which is always available as the top-1 prediction score except for inputs for which the attack is successful. The setting of targeted attacks is thus much more interesting since the target class may not be included in the top-*k* scores. As an illustration of adapting an attack to this setting, we adapt the Square Attack to the top-*k* targeted attack setting. We call this attack Square: top-*k*.

The top-*k* version of the NES attack (NES: top-*k*) modifies the original version that operates with complete prediction vector by starting from a random image of the target class (instead of the original seed in the original version), and leverages estimated gradients to gradually reduce the perturbation distance with respect to the original image while still maintaining the class prediction. This way, the confidence score of the target class is guaranteed to be in the top-*k* predictions. We speculate

that this idea can also be used to adapt the state-of-the-art Square Attack [96] by starting the attack with a random image of the target class and using corresponding perturbation generation methods to generate perturbed inputs that gradually get closer to the original seed while the target class is still in the top-k predictions. However, using the same fixed threshold on the loss function to decide when to start reducing the perturbation size, as done in NES: top-k, does not work for the Square Attack and makes the attack even less ineffective. We solve this by designing a dynamic scheduler that reduces a relatively small threshold (initially 1 in our experiments) by half if the attack is not successful in finding useful perturbations with reduced size for 10 consecutive iterations, and make the attack successful in generating useful adversarial examples.

B. Transfer Attacks

We provide additional results of transfer attacks on other target models (not covered in the main paper) in Figure 5. The overall findings still support the main claims made in Section VII.

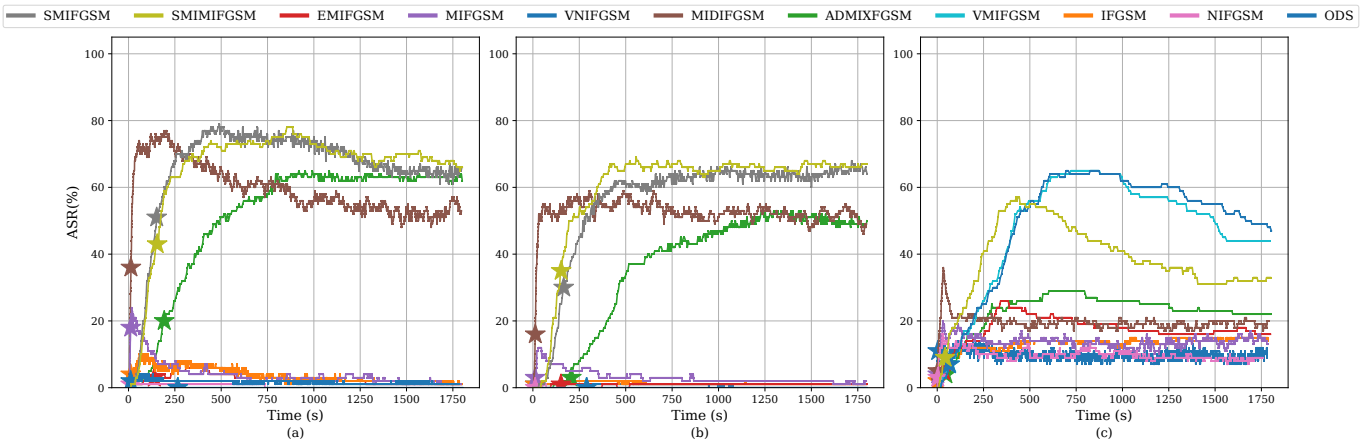

Fig. 5: ASR (y-axis) for various attacks varying across time: targeted attacks for VGG19 (a) and Resnet101 (b), and untargeted attacks for IncRes-v2$_{\text{ens}}$ (c). ASR at each iteration is computed using adversarial examples at that iteration. ASR at representative (40 for targeted, 10 for untargeted) are marked with $\star$ for each attack. Note that although SMIMI-FGSM seems to outperform other attacks in most settings, it is outperformed by VMI-FGSM and VNI-FGSM for the case of IncRes-v2$_{\text{ens}}$ (c). ASR at each iteration is computed using adversarial ex, further supporting our argument for evaluation under hard and diverse settings.

C. Query-based Attacks

Query-based attacks compare attacks by tracking ASR as queries are progressively submitted to the target model. Query cost is an important factor, as each query may incur a financial cost [124] as well as a risk of detection [206]. However, for resource constrained adversaries or situations where API costs are not a major issue (e.g., model hosted in secure enclave), the local computational cost (runtime) may be a higher priority for attackers. Adversaries in such scenarios might prefer attacks that are efficient to run locally and also require fewer queries.

In this section, we compare the bayesian optimization based attack for the hard-label setting (BayesOpt) [166] to the locally efficient RayS attack [97]. We choose these two attacks because the first attack achieves state-of-the-art performance in the low-query regime at the cost of high local runtime, while the latter achieves best performance in larger queries and is highly efficient locally. We will use these two attacks to demonstrate how the effectiveness of the attacks can change when the focus of the adversary shifts from the query cost to the local runtime cost. A secondary purpose is to check if BayesOpt attack is still the best in the low-query regime, as the BayesOpt is not compared to RayS in the original paper, despite RayS being published a year before BayesOpt at the same conference. We run untargeted attacks against Inception V3 model and set the query limit to 1,000 for the BayesOpt attack [166] and 10,000 for the RayS attack [97], all consistent with the respective original papers (we do not include targeted attacks since we could not get the BayesOpt attack to successfully generate adversarial examples in the targeted setting within the query limit).

Figure 6-a shows that the BayesOpt attack still achieves better results in the low query regime, by showing that the ASR is consistently higher than the RayS baseline. This confirms that the BayesOpt attack still achieves better performance in terms of attack success for low numbers of queries. However, when we solely measure the local runtime as the metric (Figure 6-b), the attacks proposed for the efficient attacks with sufficient queries achieve significantly higher attack success rate. Therefore, an attacker with more focus on the local cost might opt for RayS over BayesOpt in practice. Some might argue that the runtime of both attacks on a significant fraction of images are close to 0s. This is because these fraction of seeds are indeed very easy to attack and simple addition of random noise (or adding noises for a few queries) can lead to successful untargeted

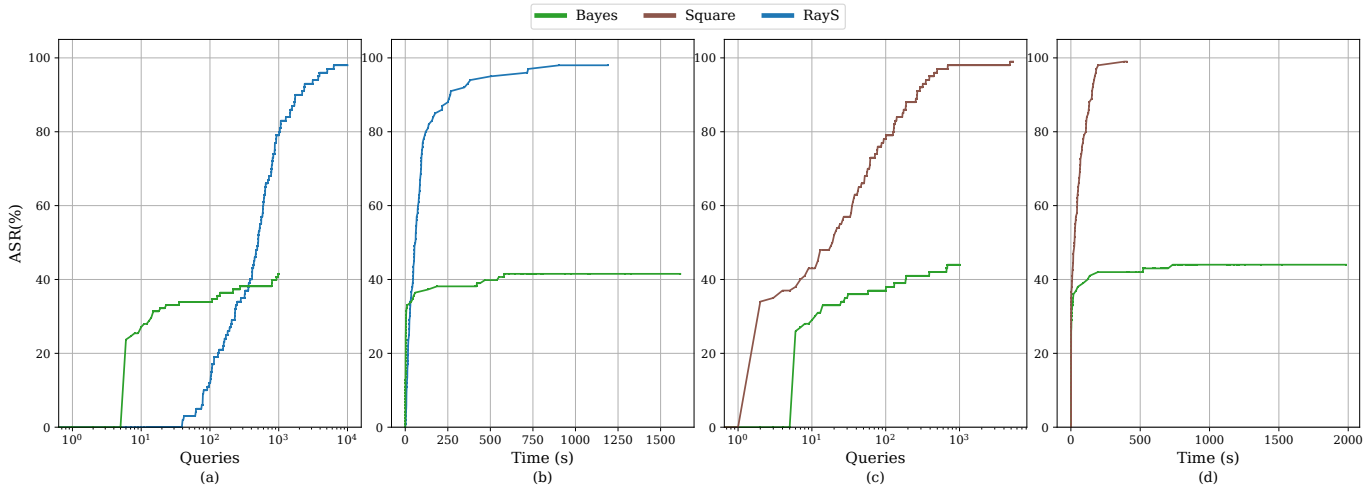

Fig. 6: ASR (y-axis) for various query-based untargeted attacks under the hard-label (a, b) and full-score setting (c, d), for Inceptionv3 target model, varying across queries (a, c) and time (b, d). ASR at each iteration is computed using adversarial examples at that number of queries.

attacks. This can also be validated by the ASR ($\approx 35\%$) of Square Attack [96] at 1 query, which adds a random noise without receiving the feedback from the target model. Different adversaries under different settings can have different priorities, such as avoiding discovery (minimizing number of queries), or wanting to be computationally efficient (minimizing local runtime). This difference in priorities, along with the demonstrated difference in attack trends, is exactly why it is important to include both kinds of metrics, instead of solely relying on the query based metric, in the future evaluation of query-based attacks.

We also repeated the same experiment in the setting of complete confidence vector, where we used the complete confidence score version of the BayesOpt attack using their corresponding implementation [166], and compared to the state-of-the-art locally efficient Square Attack [96]. We note that, there also exists another bayesian optimiation attack [120] that is reported to have even higher attack success than the results we obtained by running the BayesOpt attack above. However, the provided code by Ru et al. [120] runs extremely slowly (due to large number of CPU computations) and the attack was also not successful. The authors were also not responsive to our inquiries on possible ways to replicate their results. Therefore, we opt to use the results from the BayesOpt attack mentioned. The results are given in (Figure 6-c,d). We can see that, the locally efficient Square Attack is more efficient than the BayesOpt attack using the both metrics on the number of queries and the local runtime. This shows that, when accessing the complete confidence vector from the target, attacks explicitly proposed for the low-query regime does not seem to be the best option when compared to a more recent baseline, and encourage future research to pick the more competitive baselines for comparison.