SoK: What Have We Learned About Black-box Attacks Against Classifiers?

Anonymous Submission

***Abstract*—Dozens of papers have been written to study the vulnerability of machine learning classifiers to adversarial examples in black-box settings in which the adversary only has API access to the target classifier. These papers propose attacks for settings spanning a wide range of assumptions, making it difficult to compare the attacks and understand their effectiveness. Further, attacks are often evaluated in simplified scenarios or against weak baselines, making it difficult to discern whether proposed attacks are useful improvements in important settings. Motivated by the two observations above, we systematize the knowledge of the black-box attacks against classifiers by 1) providing a new taxonomy of black-box attacks focused on considered threat models and 2) conducting a comprehensive evaluation of representative attacks in a variety of settings. Our taxonomy reveals that although many works have been published in this space, research concentrates on a few settings while leaving others largely under-explored. Our comprehensive evaluations show that attacks that perform well in simplified settings (e.g., finding untargeted adversarial examples) often fail in other settings (e.g., targeted adversarial examples). We also evaluate a few well-performing strategies from image domains in the malware space but find that these methods often lead to worse performance, indicating that the knowledge from image classification may not easily transfer to security-relevant domains. Our systematization reveals new directions worth exploring and emphasizes the importance of evaluating attacks carefully under diverse settings.**

I. Introduction

Machine learnt models, including models using deep learning, have been found to be vulnerable to adversarial examples [52] which are specially crafted inputs designed to induce incorrect predictions from machine learning models. Most early studies of adversarial examples focused on white-box settings where the adversary is assumed to have full access to target model [52], [149], but recently there has been increasing attention paid to settings where the adversary has limited access to the target model. Such settings are a more practical threat to many deployed systems [6], [9], [63] 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 access to the model or knowledge of its parameters. Previous surveys of such attacks [10], [35], [89], [102] have categorized attacks based on their adopted methods, regardless of the disparate assumptions they make about adversarial knowledge and capabilities. These assumptions can vary wildly, however, depending on the kind of access to the model the API provides, and have a large impact on what attacks are possible. Further, proposed attacks are often evaluated on overly simplified attack settings and compared to weak baselines. This makes it unclear whether claims of out-performing prior methods hold for more challenging and practically important settings. The lack of proper and systematic evaluations poses a severe hindrance to progress in this field.

Contributions. We propose a new taxonomy for black-box attacks on classifiers organized around assumptions of threat models and apply it to attacks in the image and malware domains. Our taxonomy breaks down threat models in four dimensions (Section II): 1) the number of queries permitted to the target model, 2) the information provided by the target model's API, 3) the training data available to the adversary, and 4) the availability of similar models to the target model. We categorize the existing literature using our proposed taxonomy (Section III), and find that although there are dozens of works on several popular settings, there are many important settings that have not been well explored. We conduct a comprehensive evaluation for some of the well explored settings with competitive baselines and find that results from evaluating attacks in simplified settings do not generalize to other settings (Section IV-A and Section IV-B). We also test the transferability of techniques in image domains to malware classification (Section V). Our main findings are: 1) transfer attacks that perform well in simplified settings often do not work well in harder settings, 2) query-based attacks that are designed explicitly for limited query setting do not work well when compared to reasonable baselines, and 3) better attack methods for image classifiers do not translate well to malware. Our findings highlight the importance of comprehensive evaluations when developing attacks and the need for research focused specifically on security applications. Our taxonomy and experimental results motivate new baselines for evaluation (Section IV-C1) and directions for future work (Section III-G and Section VI).

Previous Surveys on Black-box Attacks. Several survey papers have already been written on black-box attacks in both the image and malware domains [10], [35], [89], [102], [135]. These papers all categorize attacks based on the type of methods used. Compared to these works, we first provide the taxonomy based on the threat model, which is better related to practical attack settings. These previous surveys also provide a meta-analysis on the reported results in the respective papers and identify the best performing attacks [10], [102] or reveal intriguing observations such as low transferability between different malware detection models and low evasion rate against detection [35], but all of the conclusions are drawn from results of experiments from the analyzed papers, which are done across a range of settings. Instead of drawing

conclusions based on disparate and incomparable evaluations in the source papers, we implement the representative attacks using their released source code and evaluate them in a consistent test environment under diverse attack settings. We also test the transferability of attacks from the image domain to the malware domain. The most relevant recent work is Zhao et al.'s comprehensive evaluation of transfer attacks in the image domain [199]. They focus on understanding the robustness of different defenses against the (simplified) untargeted transfer attacks at a fixed perturbation budget and comparing the visual stealthiness of different attacks with the same norm constraint. In contrast, we focus on evaluating the effectiveness of different attacks under various interesting attack settings (e.g., targeted/untargeted attacks, different perturbation budgets) and verify if attacks that perform well on easier settings also perform similarly in other settings. Further, our work covers the broader set of black-box attacks instead of just transfer attacks, providing a broader perspective on how well approaches that work on images apply to malware classification.

Relevant SoK papers. There are several previous SoK papers on adversarial machine learning, focusing on different topics ranging from categorizing attacks on audio recognition systems [1], certified robustness to adversarial examples [81] and the systematization of knowledge on general adversarial machine learning with limited coverage on black-box attacks [118]. None of these papers systematize knowledge of black-box attacks based on their applicable threat models.

II. Taxonomy for Black-Box Attacks

We propose a new taxonomy organized around the threat model assumptions of an attack using four separate dimensions for easier categorization of assumptions made by each attack. Next, we describe each dimension. Within each dimension, we describe the different categories in increasing order of adversary assumptions. In Section III, we use this taxonomy to classify black-box attacks on both image and malware classifiers. Based on analyses of these trends and the practicality of configurations, we report our insights and provide directions for future research (Section III-G).

A. Query Access

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

(a) **No Direct Access:** the adversary has absolutely no opportunity to query the target model. 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 directly query that system. This is the most challenging attack setting where the adversary has no opportunity to learn from feedback from the target model.

(b) **Limited Query Access:** the adversary can make a limited number of queries to the target model before launching the attack by submitting a generated adversarial example. This setting may arise from rate limits imposed by the target model API, a financial cost involved in making queries, or simply wanting to avoid raising suspicion.

(c) **Unlimited Access:** the most relaxed setting, where the attack is virtually unconstrained in the number of queries it can send to the target model. This may correspond to target models that do not impose rate limits, are cheap to query, or do not implement safeguards to detect suspicious activities. The most concrete example of unlimited access would be one where the adversary has access to the model on their own hardware, but it is encrypted in a secure enclave that protects its parameters.

B. API Feedback

This dimension captures the granularity of information returned by the target model's API 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 malware classifier may simply output *benign* or *malicious*.

(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, often to minimize network overhead. This setting provides more information than label-only 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) **Full-Score:** the API returns confidence scores for all classes.

C. Auxiliary Data

This dimension captures the information the adversary has about the target model's training data.

(a) **None:** the adversary has no access or knowledge of the target model's exact training data. This setting is closest to real-world APIs, where knowledge about the target model's training data is obfuscated, and often proprietary (like Open AI chatbot, etc.). However, this does not mean the adversary does not have access to *any* data at all. It may utilize domain knowledge to gather data from other sources, and given that the domain of the target model is known it is likely that the adversary can always at least obtain some data from a similar domain.

(b) **Limited Samples:** the adversary has access to some samples from the target model's exact training data.

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

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

This setting matches best with scenarios where some of the training data is public or where the adversary can poison the training data, and thus, at the very least, has knowledge of the data it contributes to the target model's training.

(c) **Sufficient Samples:** the adversary has access to a large number of samples from either the target model's exact training data, or data from a similar distribution. This setting may arise when the adversary has additional knowledge about the target model's data curation process (e.g., images found on the web). To consider the availability samples *sufficient*, instead of using an absolute number, we define it as number of representative samples adversaries can use to train models with non-trivial performance on tasks related to the task of the target model.

D. Auxiliary Models

Finally, we consider the availability of auxiliary models available to the adversary that have been trained on data similar to the target model's training data. This is a binary setting, where the value is **Yes** if the adversary somehow has access to one or more models trained on data similar to the target model's training data. This may arise in settings where the target model's training data overlaps with public data for which pre-trained models are available, or a previous iteration of the target model was released that was trained on data overlapping with its current training data.

III. Classification of Attacks

Next, based on these four proposed dimensions, we survey black-box attacks in the research literature and categorize them using the proposed taxonomy. However, it should be noted that attacks in more restrictive settings automatically apply to settings with higher levels of access. For instance, a hard-label attack for a given configuration can always be applied to the top-k or full-score setting—the adversary can simply choose to ignore the additional information of confidence scores.

The proposed taxonomy gives 3 (Query access) $\times$ 3 (API feedback) $\times$ 3 (Auxiliary data availability) $\times$ 2 (Auxiliary model availability) $=$ 54 possible configurations. However, not all of these configurations are interesting. For instance, an adversary with sufficient auxiliary data can simply train its own models on the data, and would thus not be too different from a setting where similar models are additionally available.

Table I categories attacks on image classifiers in the literature according to our taxonomy, and Table II does the same for malware classifiers. We surveyed a total of 170+ black-box attacks on classifiers in the security and machine learning literature, and based on the found attacks, we discuss the most concentrated settings below. As a background, targeted attacks aim to generate an adversarially perturbed image that is misclassified into a particular class that is different from its grountruth-class, while untargeted attacks are considered successful as long the perturbed image is misclassified into a class that is different from its grountruth class. We also note that for the malware domain, the "targeted" and 'untargeted" settings are somewhat equivalent, given most classifiers rely on either benign/malicious classifications.

A. No Similar Data, No Direct Access

This is one of the most practically feasible settings explored in the literature, and is close to real-world cyberphysical threats where malicious inputs are crafted and distributed without any feedback or additional knowledge.

1) No Similar Model: This is the most challenging attack setting as there are no reliable pretrained models to utilize. Most attacks in the image and malware domains assume the adversary has access to a sufficient number of samples from a different source domain that has no overlap with the target training data. This, in turn, allows them to generate adversarial examples on these locally trained models and then launch transfer attacks [60], [67], [70], [98], [109], [110], [112], [114], [123], [173], [188], [193]. Similar trends are observed for the malware domain [16], [99], [129]. These attacks are often termed "restricted black-box attacks" in the literature and have gained more attention from the community in recent years, especially in the image domain.

However, for images, attack performance often depends on the "similarity" between the source and target image domains, which often leads to drops in attack performance when the two domains are dissimilar, especially for targeted attacks. Very recently, attacks that only leverage the properties of images in frequency domain [194] are proposed only for the untargeted attack setting, with a similar attack based on randomized perturbations (with the utilization of domain knowledge of file formats) [72].

2) Some Similar Model: Attacks in this category are assumed to have some pretrained models on the same data as the target model, and most transfer attacks fall into this category. These attacks perform white-box attacks on the available similar models and then transfer the generated adversarial examples. These attacks mainly improve the attack success (i.e., transfer rate in this setting) by providing 1) better input transformation methods [39], [69], [88], [95], [176], [178], [179], [184], [187], [206]; 2) better gradient computation [38], [43], [45], [49]–[51], [54], [58], [69], [86], [88], [97], [150], [162], [165], [167], [171], [172], [174], [186], [201], [205]. Better gradients usually focus on uncovering correlations (e.g., spatial and temporal) hidden intra- or inter-images; 3) designing better loss functions [48], [61], [65], [66], [68], [82], [84], [94], [124], [164], [169], [171], [175], [190], [198], [203]. The loss function design gradually shifts from depending on model output to the model feature representations, as the model features contain more generic representations [73], [183], and also to focus on the more important features, typically using some interpretability techniques [133], [146]; Besides improving the optimization process of the white-box attacks, other works also focus on leveraging the similar models in more useful ways such as ensembles [87], [92] or modifying the activations of a single model in inference time [204].

Data	Model		Limited Access to Target Model			Unlimited Access to Target Model		
		No Direct Access	Hard-Label	Top-K	Full-Score	Hard-Label	Top-K	Full-Score
None	N	Frequency Manipulation [194] Train Model on Diff Source: [60], [67], [70], [98], [109], [110], [112], [114], [123], [173], [188], [193]	Bayes Optimization [139], [196]	✓	Bayes Optimization [131], [148]	Random walk: [12], [13], [40], [79], [103], [136], [145] Gradient estimation: [19], [28], [29], [93], [127], [140], [160], [197] Others: [20], [26], [106], [159], [168]	NES [63]	Gradient estimation: [2], [9], [21], [24], [42], [63], [64], [80], [85], [91], [137], [156], [195], [196] Gradient-Free: [3], [5], [22], [31], [32], [53], [105], [108], [111], [138], [153]
	Y	Input Transformation: [39], [69], [88], [95], [166], [176], [178], [179], [184], [187], [206] Gradient Manipulation: [38], [43], [45], [49]–[51], [54], [58], [69], [86], [88], [97], [150], [162], [165], [167], [171], [172], [174], [186], [201], [205] Better Loss Function: [48], [61], [65], [66], [68], [82], [84], [94], [124], [164], [169], [171], [175], [190], [198], [203] Better Local Models: [87], [92], [204]	†	†	†	✓	✓	Hybrid Attacks: [14], [25], [27], [30], [55], [62], [96], [101], [128], [147], [151], [180] Select Better surrogate: [134]
Limited	N	Train (Shallow) Models: [83], [144]	†	†	†	Surrogate Training: [116], [117], [121], [202] Better Subspace: [163]	✓	Better Subspace [163]
	Y	Random Patch [166]	†	†	†	✓	✓	✓
Sufficient	N	Train Better Local Models: [36], [132], [142], [157], [189] Training Generator: [8], [11], [57], [70], [113], [122] Input Transformation Network: [176], [184], [187]	✓	✓	✓	Optimize Gradient Estimation: [78], [191]	✓	Optimize Existing Attacks with Trained Generator: [7], [47], [107], [155], [182]
	Y	Train Auxiliary Model/ Generator: [65], [66], [70]	✓	✓	✓	✓	✓	Train Generator for Perturbation: [41], [100], [177], [181]

TABLE I: Categorization of attacks on image classifiers using our taxonomy. The first two columns correspond to availability of auxiliary data and models respectively. The remaining columns distinguish threat models based on the amount 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. For unexplored attack settings, '†' means the attack is highly relevant in practice, and "✓" denotes attack categories where we believe attacks are feasible to implement, but they are currently missing from literature.

Data	Model	Limited Access to Target Model			Unlimited Access to Target Model	
		No Direct Access	Hard-Label	Full-Score	Hard-Label	Full-Score
None	N	Randomization [72] Transfer [16], [99], [129]	Randomization [141] Genetic [130]	Genetic [130]	Randomization [18] GAN [200] Genetic [15], [34], [161] RL [23], [44], [76]	Genetic [34], [161] Explainability [129] Misc. (Hill Climb) [99]
	Y	Transfer [15], [119]	Genetic [130]	Genetic [130]	✓	
Limited	N	†	†	†	GAN [71]	✓
	Y	†	†	†	✓	✓
Sufficient	N	✓	✓	✓	RL [46], [192] Genetic [170] GAN [120], [185] Generative [59]	✓
	Y	Transfer [35], [72] [99], [158]	✓	✓	✓	✓

TABLE II: Categorization of attacks on malware classifiers using our taxonomy. Since most malware classifiers return binary predictions (benign or malicious), the top-k setting is subsumed by the full score setting.

Similarly, attacks in the malware domain also utilize transfer by improving input representation/transformations (generative adversarial networks (GANs) on top of malware bytes as images [119]) or simply having access to a variety of substitute models [15].

B. No Similar Data, Limited Feedback

Interestingly, this configuration does not seem to include the top-k setting for the image domain, despite being closest to real-world image-classification APIs. Attacks for both the hard-label and full-score settings are based on Bayesian optimization techniques [131], [139], [148], [196] while incorporating the dimensionality reduction trick since Bayesian optimization is a general-purpose optimization technique that works only with small dimensions ≈ 50. For malware, we found even fewer attacks explicitly designed for the limited-query setting, relying on either guided randomization [141] or genetic algorithms [130].

C. No Similar Data, Unlimited Feedback

This setting corresponds to most of the query-based attacks in the existing literature. We first discuss attacks that do not require similar models and divide them based on feedback details into "full-score" and "hard-label" attacks. Interestingly, only the NES [63] attack considers the top-K setting (for the image domain). After that, we discuss attacks that require access to similar models and denote these as "Hybrid Attacks". The malware domain currently lacks attacks for the 'hybrid attack' configuration.

1) Full-score attacks: These attacks can be further categorized into *gradient estimation* and *gradient-free* attacks for the image domain. The gradient estimation attacks submit queries to the target model to estimate the gradient of the target model and use the estimated gradient to perform white-box attacks [24]. Therefore, improving estimation quality with fewer queries is the main focus of these types of attacks [2], [9], [21], [42], [63], [64], [80], [85], [91], [137], [156], [195], [196]. The gradient-free attack, as the name suggests, does not rely on estimating the target model gradients. These attacks are diverse in terms of their methodologies, ranging from classical black-box optimization techniques (e.g., genetic algorithms, evolution strategies) [3], [22], [105] to efficient random search strategies [5], [31], [32], [53], [108], [111], [138], [153]. A key to the success of these attacks is to find an effective low-dimensional subspace to generate either the gradient estimation or perturbations directly. Another observation is, for attackers that aim to generate adversarial examples that satisfy the norm constraint (usually in ℓ_∞) instead of minimizing the norm, the random search-based attacks [5] perform better than the gradient estimation-based attacks.

In contrast to the image domain, attacks in the malware domain mostly rely on gradient-free attacks like genetic algorithms [34], [161] or heuristics like hill-climbing algorithms [99], since the space of features usable for training models (call graphs, API list, bytes, etc.) is much larger than the image space, which is limited to pixels. One exception uses black-box interpretability tools that estimate gradient-related statistics [129].

2) Hard-Label Attacks: For the image domain, many hard-label attacks follow the paradigm of starting from relatively large perturbations: restricting the adversarial examples to lie near the decision boundary while remaining adversarial, and then gradually reducing the perturbation size by refining the perturbations. Different methods for generating the perturbations sampled from low-dimensional space are proposed to improve query efficiency. The first types of methods are based on random walk strategy with the various sampling distributions [12], [13], [40], [79], [136], [145] or directions based on the geometry of decision boundary [103]. Another popular type of method is based on estimating the gradients, with various techniques to improve the estimation quality [19], [28], [29], [93], [127], [140], [160], [197]. There are also recent works that do not fall into the above paradigm and are shown to have the state-of-the-art performance, with diverse techniques such as random search [20], [26], [106], evolution strategies [159] or utilizing geometric properties of the decision boundary [168]. Similar to the full-score case, for norm-constrained adversaries, especially for ℓ_∞-norm, the random search-based methods are generally better than methods based on gradient estimation.

Attacks in the malware domain almost exclusively treat the target model as a plug-in in existing frameworks: reward function for reinforcement learning [23], [44], [76] and genetic algorithms [15], [34], [161], discriminator for GANs [200], and simply providing feedback for randomization techniques [18] that iteratively perturb malware until successful evasion is achieved.

3) Hybrid Attacks: These attacks have access to similar models and can also submit queries to the target model. However, existing works design attacks still based on the threat model of unlimited query access. We name these attacks "hybrid attacks" to distinguish them from pure transfer attacks or query-based attacks. There are mainly two types of hybrid attacks. The first type is to use information from similar models to help boost query-based attacks by providing better starting points (i.e., warm starting) [147] or providing better sampling space of perturbation [30], [55], [62], [96], [101], [151], [180] for the query-based attacks. The second type of attacks improve the available *similar models* with labeled queries from the target model, including fine-tuning the models [25], [27], [147], [180] or finding proper weights for individual models in the model ensemble [14], so that the transferability from these similar models will be significantly improved in the later stage. The only exception is that queries from the target model can also be combined with local explanation techniques [128] to select the most transferable single model from the set of classifiers [134].

D. Limited Samples

Knowledge of limited samples corresponds to real-world threat models, like one where the adversary participates in the

victim's training process (via federation or data upload). Despite the practicality of this setting, very few works specifically study or design attacks under this constraint.

In the image domain, when the attacker has no direct access to the victim model and also some similar models, the existing works mainly train GAN [83] or shallow layers of deep classifiers [144] to generate the adversarial examples. However, these attacks are only demonstrated for the untargeted attack setting, and investigation into the more interesting targeted setting is needed. When there are some similar models available, some attacks augment the test image (to be attacked) by patching it with random patches from images in different classes (i.e., input transformation) and then perform the standard transfer attacks [166], which is underutilizing the available information.

For the setting where attackers have unconstrained query access to the victim model, one approach is to expand the limited set of samples by submitting informative samples and also train some useful substitute models along the way [116], [117], [121] or GANs [202]. However, these attacks focus on low-resolution image datasets like MNIST or CIFAR10, leaving larger and more complex datasets like ImageNet untouched. Alternatively, the adversary may leverage available samples to learn better sampling subspace, which can benefit some full-score and hard-label query-based attacks [163]. However, these samples are directly used without labels; therefore, attackers still do not fully utilize the possible queries from the victim.

Attacks in the malware domain are even more limited, with only one work spanning the setting of unlimited query access [71], which proposes improvements in the GAN setup [71] by utilizing a small sample of clean samples from the victim's training data.

E. Sufficient Samples, No Feedback

Compared to no/limited access to samples from the target model's training dataset, this setting is much less practical, primarily because attackers under this assumption can rely on better attack transferability by training substitute models with available samples.

In these settings, even when some similar models are unavailable at first hand, the adversaries can generate better local models that have higher transferability to the target by adopting better training strategies such as standard adversarial training [36], [132], [157], only using small perturbations [142] or early stopping [189]. In addition to training classifiers, other works also focus on training generator models to produce the adversarial examples directly [8], [11], [57], [70], [113], [122], or training input transformation networks that provide better input transformation strategies to improve the standard iterative optimization-based transfer attacks [176], [184], [187]. When the adversaries also have access to the similar models, an additional auxiliary model that captures the feature distribution of the images from a particular class can be trained, which helps to generate targeted adversarial examples with higher transferability [65], [66], [70].

In this setting, attacks in the malware domain focus exclusively on transfer attacks [35], [72], [99], [158].

F. Sufficient Samples, Limited/Unlimited Feedback

When there are no similar models available, with sufficiently many samples (although may not be sufficient enough to train classifiers with state-of-the-art performance), autoencoders are trained to capture effective low-dimensional subspace, which can help to generate more query-efficient gradient estimation for the hard-label query-based attacks [78], [191] to full-score attacks [155]. Some generator models can also be trained in this setting to capture the small adversarial regions, where the random samples from the region will be adversarial [7], [47], [107], or automatically search the better hyperparameters for the state-of-the-art Square-attack [182]. When there are also similar models available, and full-prediction scores are returned from the victim model, attackers can train generator models with GAN framework [177], augmented training data with random mixing and soft labels from the target model [181], or some meta-models trained on prediction patterns of the similar models to produce generalizable perturbation patterns [41], [100].

The majority of attacks in the malware domain utilize techniques from other restricted settings like RL [46], [192], GANs [120], [185], and other methods [59], [170], with the additional benefit of having sufficient data to train substitute models or relying on a better generalization of perturbations arising from data collected from the same vendor(s).

G. Insights

From the taxonomy given in Tables I and II (for the image and malware domains respectively), we observe that most works make strong assumptions in one or more dimensions, leaving large spaces of unexplored or underexplored attack settings with moderate assumptions. Many of the most critical and practical settings have not been considered. Below, we summarize some of the insights obtained from our taxonomy that motivate the experiments in this paper's next section and suggest future work directions.

1) Under-exploration of Limited Samples: attacks mainly cover the extreme case of either having complete access to the training data of the target model or having no access to any training samples. However, in practice, it is more likely that adversaries have a limited amount of data sampled from the same distribution as the target model's training data. By leveraging this handful of samples, there may be opportunities to design stronger attacks corresponding to the other two dimensions. For example, when the attacker has (limited) query access to the target model, then querying the available samples will obtain labels from the target model, which can potentially help obtain more informative subsampling space for the query-based attacks. If there are pretrained models available, it is also feasible to take advantage of the limited available samples to generate high-value queries to the target model that reveal useful information, which is then related to model stealing attacks [115] and active learning [17].

2) Lack of Attacks in Limited Queries: in many practical settings, attackers can submit a limited number of queries to the target model API. However, only a few dedicated attacks exist for the limited query setting. In the image domain, all of these attacks adopt Bayesian optimization techniques; in the malware domain, they use genetic algorithms. In both cases, these are general-purpose optimization techniques, although equipped with some domain knowledge (e.g., sampling from a low-dimensional space). Also, in the image domain, these dedicated attacks are only evaluated against some relatively outdated baselines. As a preliminary experiment (Section IV-B), we compared the Bayesian optimization techniques and the latest advances in the unlimited query setting to check whether they are still the state-of-the-art, especially in interesting targeted attack settings. We find that they are all very ineffective in the interesting targeted settings. We can only design effective attacks in the limited query setting with the assumption of having access to a few similar models. Our taxonomy and the experimental results show the need to study attacks in the limited query setting.

3) Lack of Attacks in Top-K Predictions: Many commercial image classification APIs provide the top-k predictions. However, the only attack in the literature to consider this setting is the NES attack [63], which assumes the unrealistic unlimited query setting. Therefore, dedicated stronger attacks leveraging top-k information, ideally in the limited query setting, should be developed. One promising direction is to adapt attacks from well-explored settings (e.g., full-score or hard-label attacks). As one demonstration of this, we adapt the Square-attack of the full-score setting into the top-k setting using the main adaptation idea presented in NES. The resulting attack is shown to outperform the NES attack significantly but still performs poorly in the limited query setting (Section IV-C).

IV. Evaluation of Attacks on Image Classifiers

In this section, we first conduct experiments on transfer attacks and query-based attacks (in more practical limited query settings) in the image domain to demonstrate that attacks that perform well in simpler settings do not perform similarly in other settings. Therefore, future attacks should perform the evaluations more carefully (Section IV-A, Section IV-B). Then, inspired by our new taxonomy, we design two new attacks to act as stronger baselines for future evaluations (Section IV-C).

Since enumerating all possible combinations of attack settings is not computationally feasible, we choose to compare different attacks under the ℓ_∞ norm for the ImageNet dataset, which is the most commonly used criteria in transfer-based and query-based attacks in the image domain.

A. No Direct Access/Some Similar Models

Since the attacker has no direct access, the target model's feedback details are unnecessary. This category corresponds to the most well-studied transfer attacks in the literature [38], [39], [84], [125], [165], [166], [171], [174], [186], [198]. A recurring issue in evaluating transfer attacks is comparing attack performance with a handful of outdated baselines. Additionally, most attacks only focus on evaluating the transferability of untargeted attacks, which tend to be fairly easy but are unlikely to correspond to a security goal, as the misclassification often happens to be semantically closer classes to humans [154]. For example, as shown in Table III, using an ensemble of local models, the average untargeted transfer rate against normally trained target models is $> 99\%$ at the common perturbation budget of $\epsilon = 16/255$ in ℓ_∞-norm.

Experimental Setup. We evaluate attacks that combine the different components of 1) input transformations (4 different ones and these can be used together by stacking, and therefore there can be $C_4^1 + C_4^2 + C_4^3 + C_4^4 = 15$ combinations), and 2) 7 gradient computations (7 different ones) (Section III). To make the total number of experiments manageable, we choose the most commonly used cross-entropy loss when performing the attacks. Enumerating all these possible combinations will still leave us $15 \times 7 = 105$ attacks, which is still not manageable as we are also experimenting with other diverse and interesting attack settings (described next). Therefore, we further filter these attacks by choosing recently proposed attacks (based on reported results in the paper) that perform well (left with 26 attacks), and also additionally proposed 5 new combinations that we believe might perform well and finally end up with a total of 31 attacks. Although this is still far from an exhaustive trial of all proposed transfer attacks, it is comprehensive enough to check whether the attack with state-of-the-art performance in one setting remains as effective for all attack settings considered.

All attacks are evaluated against both standard (ResNet-101, DenseNet-201, VGG-19, Inception-v3) and robust models (IncRes-v2$_{\text{ens}}$, Inc-v3$_{\text{adv}}$), under three different perturbation budgets of $\epsilon = 4, 8, 16$ (out of 255), for both untargeted attacks and two types of targeted attacks — using a random target class (*random*) and targeting the class whose confidence score is lowest for the victim image (*hardest*). These give us a total of $3 \times 3 \times 6 \times 31 = 1674$ different attack settings. For all attack settings, the pretrained models used to generate local adversarial examples are the ensemble of VGG-16, ResNet-50, DenseNet-121, and Inception-v4. We use an ensemble of four models of different architectures since we assume an adversary who does not know that target model architecture but is aware of the popular model architectures that are likely to be used by the victim. This strategy of using an ensemble of popular model architectures empirically gives the best transfer results, compared to preliminary results of single models in our experiments, which is also consistent with the literature [38], [39], [92], [147]. Following the common setup in transfer attacks, we run the untargeted white-box optimization attacks for 10 iterations and targeted attacks for 100 iterations. Untargeted attacks are evaluated on 1,000 images, while targeted attacks are evaluated on 100 images due to the significantly increased computation time. We use the I-FGSM [75] as the white-box attack.

Results. Table III summarizes our results, where the reported attacks perform best in at least one attack setting (4 unique

Target Models	Attack	ϵ=4/255			ϵ=8/255			ϵ=16/255		
		Untargeted	Random	Hardest	Untargeted	Random	Hardest	Untargeted	Random	Hardest
ResNet-101	SMI	**89.9**	**10**	0	**98.4**	**29**	2	**99.7**	27	10
	MI-DI	76.8	**7**	1	**94.0**	**31**	**6**	**99.3**	47	**21**
	MI-DTA	62.7	5	0	82.4	**28**	**6**	**99.7**	**55**	**23**
	VNI-DTA* **(Ours)**	69.0	0	0	**95.8**	0	0	**99.9**	0	0
VGG-19	SMI	**99.9**	**19**	0	**100**	**42**	**14**	**99.2**	41	23
	MI-DI	85.4	**21**	3	**99.6**	**46**	**17**	**99.2**	**66**	**36**
	MI-DTA	68.7	15	1	**99.8**	**41**	**15**	**99.6**	58	**36**
	VNI-DTA* **(Ours)**	89.4	0	0	**99.8**	0	0	**99.6**	0	0
DenseNet-201	SMI	**90.3**	**10**	0	**98.6**	**28**	**9**	**99.8**	35	24
	MI-DI	78.4	**7**	2	**93.8**	**30**	**9**	**99.3**	59	42
	MI-DTA	65.1	3	1	**94.3**	**33**	**14**	**99.9**	**69**	**50**
	VNI-DTA* **(Ours)**	72.9	0	0	**98.1**	0	0	**99.6**	0	0
Inception-v3	SMI	**68.1**	2	1	**90.6**	**7**	5	**99.0**	15	12
	MI-DI	47.6	1	0	77.3	4	3	93.4	15	14
	MI-DTA	47.4	2	0	**89.0**	**7**	**8**	**98.0**	**30**	**32**
	VNI-DTA* **(Ours)**	49.6	0	0	**88.4**	0	0	**99.0**	0	0
IncRes-v2$_{ens}$	SMI	**23.9**	0	0	36.7	0	0	57.4	0	0
	MI-DI	13.5	0	0	19.3	0	0	34.7	0	0
	MI-DTA	17.8	0	0	36.3	0	0	67.3	2	1
	VNI-DTA* **(Ours)**	**23.1**	0	0	**56.0**	0	0	**89.0**	0	0
Inc-v3$_{adv}$	SMI	**36.8**	0	0	47.1	0	0	57.0	0	1
	MI-DI	22.9	0	0	33.8	0	0	41.3	0	0
	MI-DTA	29.1	0	0	46.4	0	0	72.1	0	1
	VNI-DTA* **(Ours)**	**36.0**	0	0	**64.3**	0	0	**86.4**	0	0

TABLE III: Evaluation of attack success rates (%) of different transfer attacks under comprehensive attack settings for varying perturbation budgets (ϵ). Apart from the basic momentum method SMI [162], the attacks mentioned above are combinations of gradient manipulation and input transformations, and are listed in the format '{Gradient Manipulation}-{Input Transformation}'. For gradient manipulation, MI [38] denotes the basic momentum method, and VNI [165] denotes variance tuned momentum method. For input transformations, DI [178] denotes the basic diversified input transformation, and DTA is the composition of DI [178], TI [39], and Admix [166] methods. Untargeted attacks are evaluated on 1,000 images while targeted attacks are evaluated on 100 images due to longer computation time. Attacks marked with $*$ are our proposed attacks by combining the input transformation and gradient manipulation. For each setting and model, we embolden attack success rates as long as 1) they are $> 5\%$, and 2) at most 5% lower than the highest attack success rate for that setting and model.

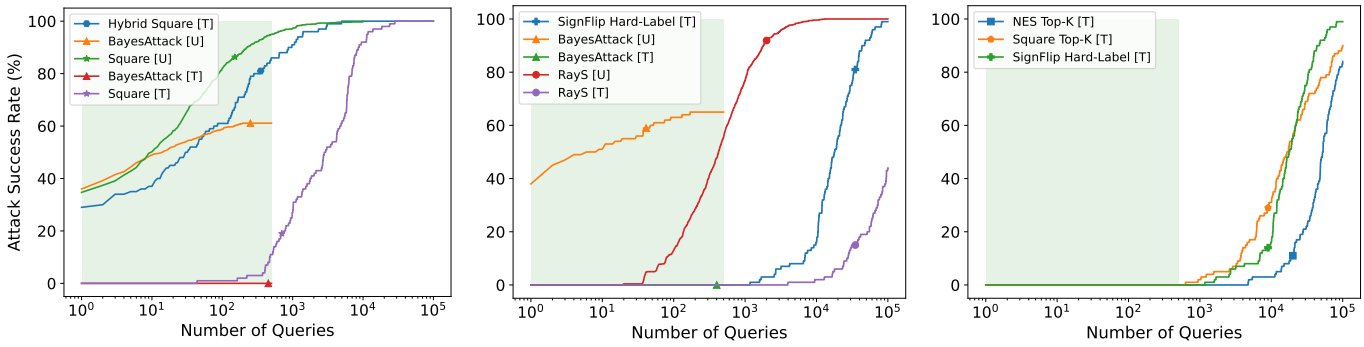
attacks in total). We note that results reported in the literature cover only a fraction of all possible transfer attacks enabled by the three components of the transfer attacks (with access to similar models) mentioned in Section III. The best-performing attacks across different attacks settings (e.g., untargeted vs. targeted, different values of ϵ for perturbation) all differ, often with a large gap between the best and second-best performing attacks, showing the fragility of transfer attack experiments and the importance of comprehensive evaluation. Evaluation of a new attack intended to be general purpose should consider all possible cases: different attacker objectives and perturbation budgets, at the very least. Some conclusions made from our results are: Regarding ϵ, for smaller values such as 4, regardless of the attack target, simple input transformation methods (e.g., no input transformation or simple transformation such as diversified input (DI) [178]) are preferred while for larger epsilon such as 16, more complex input transformations (e.g., DTA, which composes the DI [178], translation invariant (TI) [39] and Admix [166] methods) are preferred. Interestingly at the intermediate perturbation of $\epsilon = 8$, the results are also the mixture of the best attacks in $\epsilon = 4$ and $\epsilon = 16$, and the trends become noisier. One observation is, for untargeted attacks at $\epsilon = 8$, the best attacks for normally trained models are the ones that perform best at $\epsilon = 4$ while for robust models, the best attacks are ones that perform best at $\epsilon = 16$. Regarding the attack objectives, targeted attacks prefer simpler gradient manipulation, while untargeted attack usually benefits more from complicated gradient manipulation such as VNI [165].

Takeaway. Our results indicate that evaluating attacks in some simplified settings, such as higher ϵ or untargeted attacks, sometimes can be uninformative, as the results may change completely when evaluated on other more interesting (but harder) settings such as smaller ϵ or targeted settings.

B. No Data/No Similar Models/Limited Queries

Next, we consider the setting where the adversary can make a limited number of queries to the target model under APIs that provide feedback at the level of *Full Scores* and *Hard Label*, which are well-explored in the literature.

This attack category corresponds to the more practical and challenging one when attackers can interact with the target model with API queries. Therefore, advances in attacks in



(a) Full-Score Attacks (b) Hard-Label Attacks (c) Top-K Attacks

Fig. 1: Comparison of attack success rates (%) for state-of-the-art attacks across the full spectrum of queries. Limited query setting is shaded area (green) in the figures. Attacks are listed in the legends in the form 'Attack Name [U/T]', where 'U' and 'T' correspond to untargeted and targeted attacks respectively. BayesAttack has two versions designed for the settings of full-score and hard-label.

these settings are more likely to represent real threats to deployed models. It is worth noting that attacks proposed for the general unlimited query setting can also be applied to this category. At the same time, they may be outperformed by attacks particularly designed for the limited query setting. All current attacks in the literature that are designed for limited query settings are based on Bayesian optimization [131], [139], [148]. However, these Bayesian optimization-based attacks do not include comparisons with the most competent baselines (at the time of their submission) [131], [139]. Hence, we conduct comprehensive experiments to evaluate whether Bayesian optimization-based attacks are still the best option in limited query settings.

Here, "Limited" is a vague term, and queries less than 1,000 [139] and 2,000 [131] in the respective papers are assumed to be "Limited Queries". However, when we run experiments, our GPU machine becomes out of memory when the number of queries exceeds 500 (due to the increased kernel size of the Gaussian process). Therefore, we terminate the attack at 500 queries for consistency. In practice, the query limit should be posed by the target model or the financial cost of the attacker. Although our focus is on limited queries, attacks proposed for the unlimited setting can run efficiently for more queries and we also report their performance on up to 100,000 queries to get a rough idea of the performance on higher queries. We present results for each level of feedback, *hard-label* and *full scores* provided by the target model's API. In all experiments, we test both the targeted (randomly choosing the target class) and untargeted attacks against the normal Inception-v3 model at the perturbation budget of 16/255 in ℓ_∞-norm, to be consistent with the transfer attack experiments.

1) Full Scores: We pick the state-of-the-art Square-attack [5] as the representative attack in unlimited query regime and the BayesOpt attack (designed exclusively for full-score setting) [131] and full-score version of BayesAttack [139] as the particularly designed attacks in limited query setting. We note that BayesAttack is mainly designed for the hard-label setting, but the implementation also contains a version that handles full-score cases. We included it because the former BayesOpt attack is extremely slow to run ($\approx$ 100 hours to attack a single image, a similar issue is raised in the original codebase, but no response from the authors), and the authors also did not respond to our inquires. We failed to attack many images to obtain meaningful results. Therefore, for the BayesOpt attack, we directly use the results in the original paper while running both the BayesAttack and Square-attack on our machine.

Adapting to Targeted Setting. Source code of the BayesAttack does not contain targeted implementation in both their hard-label and full-score versions, although hard-label targeted results on MNIST and CIFAR10 datasets (but not ImageNet) are reported in the paper. The authors also did not respond to our inquiries about the possible adaptations to targeted settings, but the paper mentions that adaptation to hard-label targeted settings is straightforward. The loss function for the untargeted full-score attack is defined as the difference of the log of the probability of ground truth class and the log of maximum probability of other classes (i.e., equivalent to logit difference) and is minimized. We change the loss in the targeted setting by defining the loss as the (equivalent) logit difference between the target class and the maximum from other classes and then maximize the loss function. This loss design is similar to the loss function used in the BayesOpt attack.

Results. Figure 1a compares the performance of Square-attack and BayesAttack under the untargeted/targeted settings. The figure shows that in the untargeted setting, the square-attack is better than the BayesAttack, even in the limited query settings. Another observation is both attacks have reasonably high success even at 1 query, which is expected because, for ImageNet, adding some random noise (what Square-attack did with 1 query) can cause misclassification and, therefore, is

not very interesting from the security point of view. In the targeted setting, the BayesAttack is completely ineffective, while Square-attack gradually starts to achieve high success significantly after the limit of 500 queries. Of course, these may not be sufficient to conclude that Bayesian optimization-based attacks underperform other newly proposed attacks, as these modifications are mostly based on our guesses on the authors' possible implementations. Therefore, we also rerun the targeted Square-attack under the query limit of 2,000 with $\epsilon = 12.75/255$ (different from $\epsilon = 16/255$ used in Figure 1a) on 100 randomly selected images, to stay as consistent as possible with the attack settings reported in the BayesOpt paper and to compare their average statistics regarding the $Max/Median/Mean$ queries of the successful adversarial examples [131]. The original paper does not contain untargeted results and therefore we are unable to compare them in the untargeted setting. For the targeted setting, the reported attack success rate (ASR) of BayesOpt is 60% while the ASR of Square is 32%, and for BayesOpt, the average statistics are reported to be $1985/1247/1206$ (table 3 in the original paper), and the average statistics we obtained for Square-attack from experiments are $1978/1314/1029$. From these indirect[3] comparisons, in the targeted attack setting, we observe that the BayesOpt attack is probably still a better option for the full-score case with limited queries. However, more assertive claims cannot be made before comparing the two attacks under the same test environment.

2) Hard-label: In the unlimited query setting, we choose the state-of-the-art RayS attack [20] for the untargeted setting and SignFlipAttack [26] for the targeted setting, which is found to have the best performance from our experiments. Our obtained results are consistent with the numbers reported in the respective papers [20], [26]. The RayS attack is originally only designed for the untargeted setting, and we also tried to convert it into a targeted attack to see if it performs best. However, the attack success rate is inferior to the SignFlipAttack. For completeness, we also entail our conversion steps of RayS to motivate future exploration on improving the RayS attack. The Bayes attack [139] is chosen as the representative attack for the limited-query setting.

Adapting to Targeted Setting. For RayS, it works by binary searching near the decision boundary while remaining adversarial (e.g., misclassified into the target class in a targeted setting). The key is to ensure that the initial perturbation added to the original image can make the perturbed image remain adversarial. In the original paper, the perturbation $\boldsymbol{\delta}$ is initialized as $r \cdot \mathbf{1}$, where $\mathbf{1}$ is a vector of all 1s and r is a large number, which works for the untargeted setting because adding a very large perturbation can misclassify the perturbed image from the ground-truth class of the original image. However, in the targeted setting, such perturbation no longer works as the large perturbation cannot guarantee the perturbed image to be misclassified into the target class. Therefore, we change the initial perturbation to be the pixel difference between the original image $\boldsymbol{x}$ and a randomly selected image $\boldsymbol{x}_t$ from the test set that belongs to the target class (i.e., $\boldsymbol{\delta} = \boldsymbol{x_t} - \boldsymbol{x}$). Doing so can ensure that the perturbed image $\boldsymbol{x} + \boldsymbol{\delta}$ will still be in the target class in the initial round and subsequent rounds because of the steps of binary search. For the BayesAttack, in the untargeted hard-label setting, the loss function is defined as output -1 when the perturbed input still belongs to the ground-truth class of the original input and 0 when it is misclassified. We adapt to the targeted setting by simply changing the loss function to output -1 when the perturbed input is not classified into the target label and 0 when it is. In the source code, before starting the attack, the Gaussian process model of the BayesAttack is initialized by adding some random noises to the original image. We speculated that, in the hard-label setting, initializing with images near the original image may not fully capture the decision boundary near the target class, so we also tried initializing by adding random noises to a randomly selected image from the test set that belongs to the target class, and report whichever gives the better results.

Results. Figure 1b shows the performance of the three attacks under different attack settings. In the untargeted setting, we observe that the BayesAttack did outperform the RayS attack in the limited regime of ≤ 500 queries, but the gap begins shrinking drastically after 30 queries. Another observation is that the BayesAttack has a close to 40% success rate (similar to the full-score case) using 1 query compared to 0% of RayS. This is still because of the "vulnerability" of ImageNet models to random noises in untargeted settings, while RayS attack did not have the operation of adding random noise in the attack process and hence is ineffective initially. Therefore, a simple way to improve the performance of RayS attack at the initial query is first to add random noise to every sample and only run the RayS attack on the remaining unsuccessful original seeds. In the targeted setting, both the BayesAttack and SignFlipAttack attacks are ineffective at the limited query regime (< 500 queries), while the SignFlipAttack starts to achieve higher success as we allow more queries. Comparing our adapted targeted RayS and SignFlipAttack, RayS underperforms SignFlipAttack, although it performs better in untargeted settings, and we provide the possible reasons for the underperformance of RayS below to motivate future exploration.

Possible Improvements to RayS. We believe there is still room for improving the RayS attack in the targeted setting. This is because efficient attacks for a fixed ℓ_∞-norm constraint usually add perturbation of $\{-\epsilon, +\epsilon\}$ to each dimension (i.e., $\|\boldsymbol{\delta}\|_\infty = \epsilon$) [5], [20], which can be interpreted as fully exploiting the perturbation budget ϵ. However, our current way of initializing $\boldsymbol{\delta}$ in the targeted setting can lead to smaller perturbation (than ϵ) in some dimensions due to similar pixel values of $\boldsymbol{x}$ and $\boldsymbol{x}_t$ on these dimensions. These smaller perturbations cannot be increased in the subsequent attack iterations

[3]Indirect in the sense that we possibly have used different sets of images to obtain the aggregate statistics, and the target models used in these two comparisons are Inception-v3 in PyTorch and TensorFlow, which possibly can also be different.

due to the binary searching process. Therefore, we are actually under-exploiting the perturbation budget ϵ, and this may have led to inferior performance compared to the SignFlipAttack. Therefore, we may obtain stronger RayS targeted attack by generating some initial perturbation $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty = \epsilon$ that can ensure the perturbed sample is classified into the target class, and leave these explorations as future work. Although there are possibly better ways to improve these attacks, our results indicate that claims of state-of-the-art results requires careful comparison with the most competent baselines.

Takeaway. The results again highlight the importance of evaluating attacks in more diverse settings. The success in the untargeted setting is not necessarily generalizable to other settings (e.g., comparison between RayS and SignFlipAttack, SignFlipAttack and BayesAttack).

C. Proposing New Attacks

Our taxonomy indicates that the Top-K setting is largely underexplored, despite being closer to what the real-world models reveal. Also, by checking results from Section IV-B in the limited query setting, we find that state-of-the-art query-based attacks are almost ineffective in the limited query setting. In contrast, transfer attacks usually have some non-negligible success, thanks to the availability of similar models. Driven by these observations, we propose a new attack in the under-explored Top-K setting, which outperforms the existing baselines and can be used as stronger baselines for future evaluations (Section IV-C1). Then, we design new attacks by combining the best targeted transfer attacks in Section IV-A (for the Inc-3 target model) and the Square-attack to show that very effective attacks can exist for the limited query settings when some similar models are made available (Section IV-C2)

1) Top-K: We clarify that untargeted full-score attacks can be directly applied in the top-K setting and therefore is not interesting because most of these full-score attacks require the prediction confidence of the ground-truth class, which is always the top-1 prediction score and is always available to the attacker, and only becomes unavailable when the attack is successful. Therefore, for the Top-K setting, we only consider targeted attacks as the confidence score of the target class (required in the attack process) may not be contained in the top-k predictions returned from the target model during the attack process, and this makes the setting more challenging. Targeted attacks in the Top-K setting are under-explored, and currently, only the NES attack [63] explicitly considers this setting. The Top-K version of the NES attack adapts the full-score attack by starting the attack from a random image from the target class (instead of the original seed in the full-score case), and leverages estimated gradients to gradually reduce the perturbation distance with respect to the original image. This, in turn, guarantees that the confidence score of the target class is always contained in the top-k prediction. We speculate that this idea can also be used to adapt the state-of-the-art Square-attack – starting the attack from a random image from the target class and use the 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 prediction. The original NES adaptation uses a fixed threshold on the loss function to decide when to start reducing the perturbation size. However, for the Square-attack, using a fixed threshold makes the attack ineffective and we solve this by using a dynamic thresholding scheme 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 times. We did not adapt the Bayesian optimization-based attacks because of the extremely slow execution time [131] or the ineffectiveness in targeted setting [139]. For the experiments, we use the setting where only the confidence score of the top-1 prediction score is returned, consistent with the evaluation in the original NES attack [63]. Since adversaries can always ignore extra information available, we also use the results of SignFlipAttack for the hard-label setting as baselines for the relaxed Top-K setting to check whether there is a potential to improve the attack further. Since both attacks are designed for the unlimited query setting, we set the query limit to 100,000 to observe the performance of both attacks at the full spectrum of queries while focusing on the limited query setting. The result is shown in Figure 1c. From the figure, we can see that both attacks are still not very effective at the limited query regime (i.e., cutoff at 500 queries). However, the adapted Square-attack outperforms the NES attack significantly as the number of queries increases. However, the adapted Square-attack is mostly outperformed by the hard-label SignFlipAttack, which means there is some significant room for improving the attack, as the attacks are not fully utilizing the information from the target model. One possible way to improve the attack performance is to check whether existing hard-label attacks can be improved using the confidence score information, as the confidence score of the target class is now available, finer details regarding the decision boundary can be revealed for the hard-label attacks to boost their performance. Nevertheless, we still demonstrate that state-of-the-art attacks in another well-studied attack category can be adapted to design stronger attacks in the Top-K setting, and future works should focus more on this practical setting.

2) Square with Similar Models: We combine the Square-attack and the best-performing transfer attacks (for Inc-3) in the following simple way when there are some similar models available: we use all of the generated local adversarial examples from the transfer attacks as the starting points for the Square-attack, while the original attack starts by adding random noises to the original images. We name this attack Hybrid Square, and we hope these local transfers should be more directed and informative than random noises in the targeted settings. The results are reported in Figure 1a, and when we compare the performance of Hybrid Square and Square-attack, the improvement in the query efficiency is significant. For example, at the budget of 500 (the upper bound of the limited query region), Hybrid Square has attack success of 86% while the original Square-attack has only

11%. Therefore, very effective attacks are possible in the No Data/Some Similar Models/Limited Queries category, and our Hybrid Square can be a strong baseline for future evaluations.

Takeaway. For practically interesting but under-explored settings, it is possible to design better attacks by adapting from other related well-explored settings. For settings where attackers have access to more information, it is possible to design stronger attacks by combining weak baselines. Therefore, future attacks should also consider the possible existence of these (hidden) strong baselines.

V. Evaluation of Attacks on Malware Classifiers

The domain of malware samples, like most discrete domains, is inherently different from continuous domains like images, and comes with its own restrictions and domain constraints that limit the adversary's capabilities. While incorporating ℓ_p-based bounds to capture imperceptibility is sufficient for images, it is impossible for discrete inputs like malware. Unlike other discrete domains like natural language processing, where synonym-based or Hamming distance-based constraints can be loosely enforced, evasion in malware is almost purely empirical. Additionally, in their interest, adversaries are expected to craft perturbations that do not break the functionality or maliciousness of their malware– a perturbed malware that evades detection but is no longer malicious is of no use to the adversary. As apparent in Tables I and II and in our previous discussions, the image domain has received much more attention from the research community and thus has had much more profound and extensive exploration of different kinds of attacks. We are thus motivated to see if findings from the image domain can be used to improve attacks in the malware domain.

First, we utilize better gradient-based attacks in images to check if the existing basic gradient-ascent sub-routines in transfer attacks for the malware domain can be improved (Section V-A). Secondly, motivated by low transferability rates across different families of classifiers in malware [35], [90], we experiment with using ensembles of surrogate models to craft potentially stronger transfer attacks, which is demonstrated in image domain to significantly improve transferability over the single surrogate model (Section V-B).

Setup. For the following experiments, we focus on generating adversarial examples against some static malware classifiers (as the simulated black-box target models). A brief description of these models is given in Table IV. Additionally we also target a real-world black-box malware classification system – VirusTotal API[4] and report evasion rates, which we measure as the percentage of samples that can evade at least 20% of its (total of 73) detection engines. We used a sample of 100 malware samples to compute evasion rates across all our experiments.

[4]https://www.virustotal.com/gui/home/upload

Model	Dataset	Description
MalConv [4]	Ember [4]	CNN on raw bytes converted to images
GBT	Ember [4]	LightGBM classifier trained on LIEF [152] features
GBT_{ens}	Sorel [56]	Ensemble of 5 LightGBM classifiers trained with different random seeds and prediction is the average of the 5 models
$FFNN_{ens}$	Sorel [56]	Ensemble of 5 Neural Networks (NN) trained with different random seeds and prediction is the average of 5 models
FFNN	Sorel [56]	One of the 5 NNs from $FFNN_{ens}$

TABLE IV: Description of the malware classifiers used in our experiments, along with the datasets they were trained on.

A. Variations in White-Box Gradient Transfer Attacks

Both the Slack and Append [143] and Kreuk's [74] attacks utilize FGSM (in the feature embedding space) as a subroutine, with the former using only one iteration and the latter implementing the iterative version I-FGSM [75]. Given the extensive research on improvements to gradient ascent in the image domain and our findings from transfer-attacks in the image domain (Section IV-A), we propose using MI-FGSM [38] and VNI-FGSM [165] attacks instead of the I-FGSM routine. We did not incorporate the input transformation methods proposed in the image domain because those operations can easily break the malware functionality and cannot be directly applied [89]. We implement the iterative Slack-Append version attack available in secml-malware [33], and use MalConv [4] to generate perturbed malware, which are then transferred to other black-box classifiers.

Attack Variant	$FFNN_{ens}$	GBT	GBT_{ens}	VirusTotal
I-FGSM (Original)	51	9	26	**41**
MI-FGSM	51	**10**	27	40
VMI-FGSM	51	9	**28**	20

TABLE V: Evasion rates (%) for variants of the iterative Slack+Append attack, evaluated against $FFNN_{ens}$, GBT, GBT_{ens} and the VirusTotal API.

Results are reported in Table V and from the table, we can observe that, for the static malware classifiers (given in Table IV and used as the simulated black-box models), the transferability remains nearly the same. This observation is in line with prior findings in the literature that experimenting with complicated gradient-search techniques like DeepFool and C&W against image-based classifiers for malware [72] does not lead to improved transferability, suggesting a possible roadblock on possible improvements purely for better gradient search in transfer attacks for malware. Interestingly though, we observe a significant drop in performance when evaluating the local transfers against the truly black-box VirusTotal API: evasion rates go down from 41% for the original variant I-FGSM, to 40% for MI-FGSM and as low as 20% for VNI-FGSM. One possible reason for this drop could be an

increased tendency of the local perturbations overfitting to certain types of features while APIs like VirusTotal utilize multiple classifiers that look at a wide variety of features.

Takeaway. Attempts at improving gradient search in attacks for malware do not necessarily lead to improved evasion and can in fact hurt transferability significantly. Simply plugging in improved methods based on findings from different domains may not be the best approach to transfer knowledge.

B. Auxiliary Model Ensembles

Although ensembles have been proposed for target models (and are in fact used in systems like VirusTotal), none of the works surveyed for Windows PE malware utilize ensembles of surrogate models for crafting perturbations. In fact, to the best of our knowledge, looking across *all* malware sub-domains (Windows, Android, PDF, etc.), only one work proposed using local ensembles of models [77] for Android. Since this is a common approach in the image domain and is expected to help with transferability [92], we experiment with ensembles of surrogate models for the malware domain and test whether it can lead to improved transferability to the VirusTotal API. Similar to experiments in the image domain described in Section IV-A, we emulate ensembles by averaging predictions across all classifiers in the ensemble for any given input. We experiment with ensembles of models that have the same architecture but trained with different random seeds (e.g., GBT_{ens}, FFNN_{ens}) and models are from different families or use different feature processing techniques (e.g., MalConv and GBT). As for the choice of attacks, unlike the image domain, classifiers in the malware domain utilize a variety of features, ranging from raw bytes to hand-crafted features and makes it impossible to utilize them simultaneously for the white-box attacks (e.g., gradient attacks) that directly modify the extracted features. Thus, we choose to use the DOS Header Extend attack [35] to generate perturbed malwares.

The results are summarized in Table VI. For GBT and FNNN models, using an ensembles of same architecture but trained with different seeds (e.g., GBT_{ens}) is nearly as effective as ensembles of models with different feature processing techniques, and both types of ensembles are more effective than using a single surrogate model. For example, the low evasion rate of 37% for a single GBT model can be improved to near 70% using GBT_{ens}or $\text{FFNN}_{\text{ens}}$and GBT, and similarly for the case of FFNN.

However, when we also consider MalConv, we did not see improvements in the evasion rates by using the ensembles. Concretely, the best ensemble (MalConv, GBT_{ens}, FFNN_{ens}) has an evasion rate only as high as that when using just a MalConv model. This finding provides two messages: on one hand, using the ensemble is beneficial in a query-restricted setting, as an ensemble may easily contain the best surrogate model for transferability (e.g., MalConv in our experiments) and the adversary avoids having to guess the best surrogate from a choice of local auxiliary models. On the other hand, the lack of significant improvement over the best matching surrogate model suggests there scope for constructing better ensembles, either by utilizing more diverse models or better ensemble aggregation techniques. As an initial attempt to improve the simple aggregation method of averaging the predictions, we propose and experiment with the following variant: starting with a given malware sample, for each of the models (say m_i) from the set (MalConv, GBT, GBT_{ens}, FFNN_{ens}), we generate the perturbed malware sample (say x_i) individually. Then, we check the transferability of each x_i against the ensemble of the remaining models $m_{j\neq i}$, where the prediction of $m_{j\neq i}$ is still based on averaging all the predictions. Then, the x_i with the highest transferability is chosen as the final operturbed sample for the given malware sampple and is sent to the VirusTital API. However, this variant leads to a very low evasion rate (34%) and is unsuccessflul.

Ensemble of Model(s)	VirusTotal
MalConv	**71**
GBT	37
GBT_{ens}	70
FFNN	60
FFNN_{ens}	69
MalConv, GBT	69
MalConv, GBT_{ens}	70
MalConv, FFNN_{ens}	47
GBT, FFNN_{ens}	69
GBT_{ens}, FFNN_{ens}	67
MalConv, GBT, FFNN_{ens}	68
MalConv, GBT_{ens}, FFNN_{ens}	66

TABLE VI: Evasion rates (%) for the DOS Header Extend attack while using a variety of local ensemble models, evaluated against the VirusTotal API.

Takeaway. The utilization of ensembles of models has the potential of alleviating issues with the generalization of attacks between classifiers that utilize different features, like raw bytes or static-features. Nonetheless, the absence of any improvement in performance on combination compared to best matching single surrogate hints at the possibility of better techniques for collating information from auxiliary models.

VI. Discussions and New Directions

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

Many Interesting Settings Underexplored. In the image domain, our taxonomy reveals that many attack categories align well with practical applications but are not well covered in the existing literature. The particularly interesting settings are ones that allow attackers to submit a limited number of queries, only revealing hard-label or top-k prediction scores from the target model, or only having access to no or a limited number of samples that are sampled from the same distribution of the target data. Our preliminary results on the Top-K setting validate that transferring knowledge from other less realistic

but well-explored settings is indeed feasible (Section IV-C). However, there is still some significant room for improvement. Therefore, we encourage the research community to, while proposing stronger attacks in well-explored domains with multiple assumptions (corresponding to moving down/right in the taxonomy, Table I), also consider the possibility of adaptation to the more realistic attack settings.

Careful Evaluation of Future Attacks. Future attacks should be evaluated carefully before making state-of-the-art claims. We demonstrated this by first showing that, in transfer attacks, complicated transfer attacks often perform better in the simpler attack settings but become ineffective once the attack settings become harder. Second, in the attack settings where attackers have access to similar models, receive full-prediction scores, and can submit limited queries to the target, the attack can be highly successful in the limited query setting. Therefore, future attacks should do the evaluations carefully and attacks that are proposed for categories with richer information available should also consider the possible existence of stronger baselines that are enabled by the additional information.

Lessons on Transferring Knowledge Across Domains. Findings in malware suggest that improvements from image-based attacks do not necessarily translate to attacks in the malware domain. We worry that these might be true in the general sense. First, while using ensembles of auxiliary models lends the benefit of not having to guess the target model's feature processing pipeline, there is not much improvement from using ensembles once the best single local model is identified, which is a completely different observation compared to ones in image domains, where using an ensemble of models always leads to significantly improved performance [92]. As discussed in Section V, malware space has its unique constraints (e.g., preserving file format, ensuring executability, and maliciousness that is independent of the target classifier) that are drastically different from the image domain. The compatibility of certain techniques does not necessarily imply transferability, and future designs should explicitly consider the domain constraint when transferring knowledge. Another example is (state-of-the-art) gradient-free query based attacks primarily work well on images as adversarial examples in images are often assumed to reside in some low-dimensional spaces [163] and the perturbation also show some regional homogeneity [86], which makes the efficient random search feasible. At the same time, such properties are hardly true for malwares and are often dependent on the type of feature processor used.

Limitations. In this work, we only systematize the knowledge of inference time black-box attacks, which implicitly assumes the target models are static. However, real-world classification models are constantly being improved over time, mostly by fine-tuning on unverified sources from the internet that potentially are controlled by some malicious adversaries [104], [126]. When faced with adversaries that can conduct both the training and inference time attacks, the attacker can be very powerful, but our current taxonomy cannot capture the dynamically changing possibly poisoned target models [37].

REFERENCES

[1] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick 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.

[2] Abdullah Al-Dujaili and Una-May O'Reilly. Sign Bits Are All You Need for Black-Box Attacks. In *International Conference on Learning Representations*, 2019.

[3] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. GenAttack: Practical Black-box Attacks with Gradient-Free Optimization. In *The Genetic and Evolutionary Computation Conference*, 2019.

[4] Hyrum S Anderson and Phil Roth. EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. *arXiv:1804.04637*, 2018.

[5] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, 2020.

[6] Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin A. Roundy. "real attackers don't compute gradients": Bridging the gap between adversarial ML research and practice. *arXiv:2212.14315*, 2022.

[7] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. Improving Query Efficiency of Black-box Adversarial Attack. In *European Conference on Computer Vision*, 2020.

[8] Shumeet Baluja and Ian Fischer. Adversarial Transformation Networks: Learning to Generate Adversarial Examples. *arXiv:1703.09387*, 2017.

[9] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical Black-box Attacks on Deep Neural Networks using Efficient Query Mechanisms. In *European Conference on Computer Vision*, 2018.

[10] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. A Survey of Black-Box Adversarial Attacks on Computer Vision Models. *arXiv:1912.01667*, 2019.

[11] Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and Will Hamilton. Adversarial Example Game. In *Advances in Neural Information Processing Systems*, 2020.

[12] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*, 2017.

[13] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[14] Zikui Cai, Chengyu Song, Srikanth Krishnamurthy, Amit Roy-Chowdhury, and M Salman Asif. Blackbox Attacks via Surrogate Ensemble Search. *arXiv:2208.03610*, 2022.

[15] Raphael Labaca Castro, Corinna Schmitt, and Gabi Dreo. AIMED: Evolving Malware with Genetic Programming to Evade Detection. In *International Conference on Trust, Security and Privacy in Computing and Communications / International Conference on Big Data Science and Engineering*, 2019.

[16] Fabrício Ceschin, Marcus Botacin, Heitor Murilo Gomes, Luiz S Oliveira, and André Grégio. Shallow Security: on the Creation of Adversarial Variants to Evade Machine Learning-Based Malware Detectors. In *3rd Reversing and Offensive-oriented Trends Symposium*, 2019.

[17] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring Connections Between Active Learning and Model Extraction. In *USENIX Security Symposium*, 2020.

[18] Bingcai Chen, Zhongru Ren, Chao Yu, Iftikhar Hussain, and Jintao Liu. Adversarial Examples for CNN-Based Malware Detectors. *IEEE Access*, 2019.

[19] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. In *IEEE Symposium on Security and Privacy*, 2020.

[20] Jinghui Chen and Quanquan Gu. RayS: A Ray Searching Method for Hard-label Adversarial Attack. In *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.

[21] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A Frank-Wolfe Framework for Efficient and Effective Adversarial Attacks. In *AAAI conference on artificial intelligence*, 2020.

[22] Jinyin Chen, Mengmeng Su, Shijing Shen, Hui Xiong, and Haibin Zheng. POBA-GA: Perturbation Optimized Black-Box Adversarial Attacks via Genetic Algorithm. *Computers & Security*, 2019.

[23] Jun Chen, Jingfei Jiang, Rongchun Li, and Yong Dou. Generating Adversarial Examples for Static PE Malware Detector Based on Deep Reinforcement Learning. In *Journal of Physics: Conference Series*. IOP Publishing, 2020.

[24] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui 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.

[25] Sizhe Chen, Zhehao Huang, Qinghua Tao, and Xiaolin Huang. QueryNet: Attack by Multi-Identity Surrogates. *arXiv:2105.15010*, 2021.

[26] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting Decision-Based Black-Box Adversarial Attacks with Random Sign Flip. In *European Conference on Computer Vision*. Springer, 2020.

[27] Zhiyu Chen, Jianyu Ding, Fei Wu, Chi Zhang, Yiming Sun, Jing Sun, Shangdong Liu, and Yimu Ji. An Optimized Black-Box Adversarial Simulator Attack Based on Meta-Learning. *Entropy*, 2022.

[28] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-Efficient Hard-label Black-box Attack: An Optimization-based Approach. In *International Conference on Learning Representations*, 2018.

[29] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. In *International Conference on Learning Representations*, 2019.

[30] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving Black-box Adversarial Attacks with a Transfer-based Prior. In *Advances in Neural Information Processing Systems*, 2019.

[31] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks. In *AAAI Conference on Artificial Intelligence*, 2022.

[32] Francesco Croce and Matthias Hein. Sparse and Imperceivable Adversarial Attacks. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[33] Luca Demetrio and Battista Biggio. secml-malware: Pentesting Windows Malware Classifiers with Adversarial EXEmples in Python, 2021.

[34] Luca Demetrio, Battista Biggio, Giovanni Lagorio, Fabio Roli, and Alessandro Armando. Functionality-preserving Black-box Optimization of Adversarial Windows Malware. *IEEE Transactions on Information Forensics and Security*, 2021.

[35] Luca Demetrio, Scott E Coull, Battista Biggio, Giovanni Lagorio, Alessandro Armando, and Fabio Roli. Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection. *ACM Transactions on Privacy and Security*, 2021.

[36] Zhun Deng, Linjun Zhang, Kailas Vodrahalli, Kenji Kawaguchi, and James Y Zou. Adversarial Training Helps Transfer Learning via Better Representations. In *Advances in Neural Information Processing Systems*, 2021.

[37] Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust PAC learning under evasion and hybrid attacks. In *IEEE International Conference on Machine Learning and Applications*, 2020.

[38] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[39] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[40] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient Decision-based Black-box Adversarial Attacks on Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[41] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient Meta Attack to Deep Neural Networks. In *International Conference on Learning Representations*, 2019.

[42] Yali Du, Meng Fang, Jinfeng Yi, Jun Cheng, and Dacheng Tao. Towards Query Efficient Black-box Attacks: An Input-free Perspective. In *11th ACM Workshop on Artificial Intelligence and Security*, 2018.

[43] Yexin Duan, Junhua Zou, Xingyu Zhou, Wu Zhang, Zhengyun He, Dazhi Zhan, Jin Zhang, and Zhisong Pan. Adversarial Attack via Dual-Stage Network Erosion. *Computers & Security*, 2022.

[44] Mohammadreza Ebrahimi, Jason Pacheco, Weifeng Li, James Lee Hu, and Hsinchun Chen. Binary Black-Box Attacks Against Static Malware Detectors with Reinforcement Learning in Discrete Action Spaces. In *IEEE Security and Privacy Workshops*. IEEE, 2021.

[45] Shuman Fang, Jie Li, Xianming Lin, and Rongrong Ji. Learning to Learn Transferable Attack. In *AAAI Conference on Artificial Intelligence*, 2022.

[46] Zhiyang Fang, Junfeng Wang, Boya Li, Siqi Wu, Yingjie Zhou, and Haiying Huang. Evading Anti-Malware Engines With Deep Reinforcement Learning. *IEEE Access*, 2019.

[47] Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting Black-Box Attack with Partially Transferred Conditional Adversarial Distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[48] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. FDA: Feature Disruptive Attack. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[49] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise Attack for Fooling Deep Neural Network. In *European Conference on Computer Vision*. Springer, 2020.

[50] Lianli Gao, Qilong Zhang, Jingkuan Song, and Heng Tao Shen. Patch-wise++ Perturbation for Adversarial Targeted Attacks. *CoRR*, 2020.

[51] Lianli Gao, Qilong Zhang, Xiaosu Zhu, Jingkuan Song, and Heng Tao Shen. Staircase Sign Method for Boosting Adversarial Attacks. *CoRR*, 2021.

[52] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2014.

[53] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple Black-box Adversarial Attacks. In *International Conference on Machine Learning*. PMLR, 2019.

[54] Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating Linearly Improves Transferability of Adversarial Examples. In *Advances in Neural Information Processing Systems*, 2020.

[55] Yiwen Guo, Ziang Yan, and Changshui Zhang. Subspace Attack: Exploiting Promising Subspaces for Query-Efficient Black-box Attacks. In *Advances in Neural Information Processing Systems*, 2019.

[56] Richard Harang and Ethan M Rudd. Sorel-20m: A large scale benchmark dataset for malicious pe detection. *arXiv preprint arXiv:2012.07634*, 2020.

[57] Atiye Sadat Hashemi, Andreas Bär, Saeed Mozaffari, and Tim Fingscheidt. Transferable Universal Adversarial Perturbations Using Generative Models . *arXiv:2010.14919*, 2020.

[58] Zhengyun He, Yexin Duan, Wu Zhang, Junhua Zou, Zhengfang He, Yunyun Wang, and Zhisong Pan. Boosting Adversarial Attacks with Transformed Gradient. *Computers & Security*, 2022.

[59] Weiwei Hu and Ying Tan. Black-Box Attacks against RNN based Malware Detection Algorithms. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[60] Zhaoxin Huan, Yulong Wang, Xiaolu Zhang, Lin Shang, Chilin Fu, and Jun Zhou. Data-Free Adversarial Perturbations for Practical Black-Box Attack. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2020.

[61] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing Adversarial Example Transferability with an Intermediate Level Attack. In *IEEE/CVF international conference on computer vision*, 2019.

[62] Zhichao Huang and Tong Zhang. Black-Box Adversarial Attack with Transferable Model-based Embedding. *arXiv:1911.07140*, 2019.

[63] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *International Conference on Machine Learning*. PMLR, 2018.

[64] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior Convictions: Black-Box Adversarial Attacks with Bandits and Priors. In *International Conference on Learning Representations*, 2018.

[65] Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability. *Advances in Neural Information Processing Systems*, 2020.

[66] Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. Transferable Perturbations of Deep Feature Distributions. In *International Conference on Learning Representations*, 2020.

[67] Nathan Inkawhich, Kevin J Liang, Jingyang Zhang, Huanrui Yang, Hai Li, and Yiran 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.

[68] Nathan Inkawhich, Wei Wen, Hai Helen Li, and Yiran Chen. Feature Space Perturbations Yield More Transferable Adversarial Examples. In *EEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[69] Donggon Jang, Sanghyeok Son, and Dae-Shik 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.

[70] Krishna kanth Nakka and Mathieu Salzmann. Learning Transferable Adversarial Perturbations. In *Advances in Neural Information Processing Systems*, 2021.

[71] Masataka Kawai, Kaoru Ota, and Mianxing Dong. Improved MalGAN: Avoiding Malware Detector by Leaning Cleanware Features. In *International conference on artificial intelligence in information and communication*. IEEE, 2019.

[72] Aminollah Khormali, Ahmed Abusnaina, Songqing Chen, DaeHun Nyang, and Aziz Mohaisen. COPYCAT: Practical Adversarial Attacks on Visualization-Based Malware Detection. *CoRR*, 2019.

[73] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *International Conference on Machine Learning*. PMLR, 2019.

[74] Felix Kreuk, Assi Barak, Shir Aviv-Reuven, Moran Baruch, Benny Pinkas, and Joseph Keshet. Deceiving end-to-end deep learning malware detectors using adversarial examples. *CoRR*, 2018.

[75] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. In *International Conference on Learning Representations*, 2016.

[76] Raphael Labaca-Castro, Sebastian Franz, and Gabi Dreo Rodosek. AIMED-RL: Exploring Adversarial Malware Examples with Reinforcement Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021.

[77] Deqiang Li and Qianmu Li. Adversarial Deep Ensemble: Evasion Attacks and Defenses for Malware Detection. *IEEE Transactions on Information Forensics and Security*, 2020.

[78] Huichen Li, Linyi Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Nonlinear Projection Based Gradient Estimation for Query Efficient Blackbox Attacks. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

[79] Jie Li, Rongrong Ji, Peixian Chen, Baochang Zhang, Xiaopeng Hong, Ruixin Zhang, Shaoxin Li, Jilin Li, Feiyue Huang, and Yongjian Wu. Aha! Adaptive History-driven Attack for Decision-based Black-box Models. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[80] Jie Li, Rongrong Ji, Hong Liu, Jianzhuang Liu, Bineng Zhong, Cheng Deng, and Qi Tian. Projection & Probability-Driven Black-Box Attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[81] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. SoK: Certified Robustness for Deep Neural Networks. *IEEE Symposium on Security and Privacy*, 2020.

[82] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards Transferable Targeted Attack. In *IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[83] Qizhang Li, Yiwen Guo, and Hao Chen. Practical No-box Adversarial Attacks against DNNs. In *Advances in Neural Information Processing Systems*, 2020.

[84] Qizhang Li, Yiwen Guo, and Hao Chen. Yet Another Intermediate-Level Attack. In *European Conference on Computer Vision*. Springer, 2020.

[85] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing 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.

[86] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan Yuille. Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses. In *European Conference on Computer Vision*. Springer, 2020.

[87] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning Transferable Adversarial Examples via Ghost Networks. In *AAAI Conference on Artificial Intelligence*, 2020.

[88] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In *International Conference on Learning Representations*, 2019.

[89] Xiang Ling, Lingfei Wu, Jiangyu Zhang, Zhenqing Qu, Wei Deng, Xiang Chen, Chunming Wu, Shouling Ji, Tianyue Luo, Jingzheng Wu, et al. Adversarial Attacks against Windows PE Malware Detection: A Survey of the State-of-the-Art. *arXiv:2112.12310*, 2021.

[90] Hao Liu, Wenhai Sun, Nan Niu, and Boyang Wang. MultiEvasion: Evasion Attacks Against Multiple Malware Detectors. In *IEEE Conference on Communications and Network Security*, 2022.

[91] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via Zeroth-Order Oracle. In *International Conference on Learning Representations*, 2018.

[92] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into Transferable Adversarial Examples and Black-box Attacks. *International Conference on Learning Representations*, 2017.

[93] Yujia Liu, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. A geometry-inspired decision-based attack. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[94] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's Afraid of Adversarial Queries? The Impact of Image Modifications on Content-based Image Retrieval. In *International Conference on Multimedia Retrieval*, 2019.

[95] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency Domain Model Augmentation for Adversarial Attack. In *European Conference on Computer Vision*. Springer, 2022.

[96] Nicholas A Lord, Romain Mueller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. *arXiv:2203.08725*, 2022.

[97] Qiming Lu, Shikui Wei, Haoyu Chu, and Yao Zhao. Towards Transferable 3D Adversarial Attack. In *ACM Multimedia Asia*. Association for Computing Machinery, 2021.

[98] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples with Dispersion Reduction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[99] Keane Lucas, Mahmood Sharif, Lujo Bauer, Michael K Reiter, and Saurabh Shintre. Malware Makeover: Breaking ML-based Static Analysis by Modifying Executable Bytes. In *2021 ACM Asia Conference on Computer and Communications Security*, 2021.

[100] Chen Ma, Li Chen, and Jun-Hai Yong. Simulating Unknown Target Models for Query-Efficient Black-box Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[101] Chen Ma, Shuyu Cheng, Li Chen, Jun Zhu, and Junhai Yong. Switching Transferable Gradient Directions for Query-Efficient Black-Box Adversarial Attacks. *arXiv:2009.07191*, 2020.

[102] Kaleel Mahmood, Rigel Mahmood, Ethan Rathbun, and Marten Van Dijk. Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks. *IEEE Access*, 2021.

[103] Thibault Maho, Teddy Furon, and Erwan Le Merrer. SurFree: a fast surrogate-free black-box attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[104] Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. How Robust are Randomized Smoothing Based Defenses to Data Poisoning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[105] Laurent Meunier, Jamal Atif, and Olivier Teytaud. Yet another but more efficient black-box adversarial attack: tiling and evolution strategies. *arXiv:1910.02244*, 2019.

[106] Kim AB Midtlid, Johannes Åsheim, and Jingyue Li. Magnitude Adversarial Spectrum Search-based Black-box Attack against Image Classification. In *15th ACM Workshop on Artificial Intelligence and Security*, 2022.

[107] Hadi Mohaghegh Dolatabadi, Sarah Erfani, and Christopher Leckie. AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows. *Advances in Neural Information Processing Systems*, 2020.

[108] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious Black-Box Adversarial Attacks via Efficient Combinatorial Optimization. In *36th International Conference on Machine Learning*. PMLR, 2019.

[109] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable Data-free Objective for Crafting Universal Adversarial Perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[110] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. Fast Feature Fool: A data independent approach to universal adversarial perturbations. *arXiv:1707.05572*, 2017.

[111] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple Black-Box Adversarial Attacks on Deep Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[112] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-Domain Transferability of Adversarial Perturbations. *Advances in Neural Information Processing Systems*, 2019.

[113] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. On Generating Transferable Targeted Perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[114] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable Adversarial Attack based on Perceptual Metric. *arXiv preprint arXiv:1811.09020*, 2018.

[115] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff Nets: Stealing Functionality of Black-box Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[116] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv:1605.07277*, 2016.

[117] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *ACM on Asia conference on computer and communications security*, 2017.

[118] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. SoK: Security and Privacy in Machine Learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.

[119] Daniel Park, Haidar Khan, and Bülent Yener. Generation & Evaluation of Adversarial Examples for Malware Obfuscation. In *18th IEEE International Conference On Machine Learning And Applications*, 2019.

[120] Xiaowei Peng, Hequn Xian, Qian Lu, and Xiuqing Lu. Semantics aware adversarial malware examples generation for black-box attacks. *Applied Soft Computing*, 2021.

[121] Li Pengcheng, Jinfeng Yi, and Lijun Zhang. Query-Efficient Black-box Attack by Active Learning. In *IEEE International Conference on Data Mining*, 2018.

[122] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative Adversarial Perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[123] Yunxiao Qin, Yuanhao Xiong, Jinfeng Yi, and Cho-Jui Hsieh. Adversarial Attack across Datasets. *arXiv:2110.07718*, 2021.

[124] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation. In *Advances in Neural Information Processing Systems*, 2022.

[125] Zeyu Qin, Yanbo Fan, Yi Liu, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation. In *Advances in Neural Information Processing Systems*, 2022.

[126] Evani Radiya-Dixit and Florian Tramèr. Data Poisoning Won't Save You From Facial Recognition. In *International Conference on Learning Representations*, 2022.

[127] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: a geometric framework for black-box adversarial attacks. In *IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[128] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.

[129] Ishai Rosenberg, Shai Meir, Jonathan Berrebi, Ilay Gordon, Guillaume Sicard, and Eli Omid David. Generating End-to-End Adversarial Examples for Malware Classifiers Using Explainability. In *2020 International Joint Conference on Neural Networks*. IEEE, 2020.

[130] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Query-Efficient Black-Box Attack Against Sequence-Based Malware Classifiers. In *Annual Computer Security Applications Conference*, 2020.

[131] Binxin Ru, Adam Cobb, Arno Blaas, and Yarin Gal. Bayesopt Adversarial Attack. In *International Conference on Learning Representations*, 2019.

[132] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do Adversarially Robust ImageNet Models Transfer Better? *Advances in Neural Information Processing Systems*, 2020.

[133] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *IEEE international conference on computer vision*, 2017.

[134] Giorgio Severi, Will Pearce, and Alina Oprea. Bad Citrus: Reducing Adversarial Costs with Model Distances. *arXiv:2210.03239*, 2022.

[135] Ali Shafiei, Vera Rimmer, Ilias Tsingenopoulos, Lieven Desmet, and Wouter Joosen. Position Paper: On Advancing Adversarial Malware Generation Using Dynamic Features. In *Proceedings of the 1st Workshop on Robust Malware Analysis*, pages 15–20, 2022.

[136] Yucheng Shi, Yahong Han, and Qi Tian. Polishing Decision-Based Adversarial Noise With a Customized Sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[137] Yucheng Shi, Siyu Wang, and Yahong Han. Curls & Whey: Boosting Black-Box Adversarial Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

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

[139] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and Zico 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.

[140] Carl-Johann Simon-Gabriel, Noman Ahmed Sheikh, and Andreas Krause. PopSkipJump: Decision-Based Attack for Probabilistic Classifiers. In *38th International Conference on Machine Learning*. PMLR, 2021.

[141] Wei Song, Xuezixiang Li, Sadia Afroz, Deepali Garg, Dmitry Kuznetsov, and Heng Yin. Automatic Generation of Adversarial Examples for Interpreting Malware Classifiers. *arXiv:2003.03100*, 2020.

[142] Jacob Springer, Melanie Mitchell, and Garrett Kenyon. A Little Robustness Goes a Long Way: Leveraging Robust Features for Targeted Transfer Attacks. In *Advances in Neural Information Processing Systems*, 2021.

[143] Octavian Suciu, Scott E Coull, and Jeffrey Johns. Exploring adversarial examples in malware detection. In *2019 IEEE Security and Privacy Workshops*. IEEE, 2019.

[144] Chenghao Sun, Yonggang Zhang, Wan Chaoqun, Qizhou Wang, Ya Li, Tongliang Liu, Bo Han, and Xinmei Tian. Towards Lightweight Black-Box Attacks against Deep Neural Networks. *arXiv:2209.14826*, 2022.

[145] Xuxiang Sun, Gong Cheng, Lei Pei, and Junwei Han. Query-efficient decision-based attack via sampling distribution reshaping. *Pattern Recognition*, 2022.

[146] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *International conference on machine learning*. PMLR, 2017.

[147] Fnu Suya, Jianfeng Chi, David Evans, and Yuan Tian. Hybrid Batch Attacks: Finding Black-box Adversarial Examples with Limited Queries. In *29th USENIX Security Symposium*, 2020.

[148] Fnu Suya, Yuan Tian, David Evans, and Paolo Papotti. Query-limited Black-box Attacks to Classifiers. *arXiv:1712.08713*, 2017.

[149] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*, 2013.

[150] Hao Tan, Zhaoquan Gu, Le Wang, Huan Zhang, Brij B Gupta, and Zhihong Tian. Improving Adversarial Transferability by Temporal and

Spatial Momentum in Urban Speaker Recognition Systems. *Computers and Electrical Engineering*, 2022.

[151] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be Transferred: Output Diversification for White- and Black-box Attacks. In *Advances in Neural Information Processing Systems*, 2020.

[152] Romain Thomas. Lief - library to instrument executable formats. https://lief.quarkslab.com/, apr 2017.

[153] Hoang Tran, Dan Lu, and Guannan Zhang. Exploiting the Local Parabolic Landscapes of Adversarial Losses to Accelerate Black-Box Adversarial Attack. In *European Conference on Computer Vision*. Springer, 2022.

[154] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, 2020.

[155] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks. In *AAAI Conference on Artificial Intelligence*, 2019.

[156] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In *International Conference on Machine Learning*. PMLR, 2018.

[157] Francisco Utrera, Evan Kravitz, N Benjamin Erichson, Rajiv Khanna, and Michael W Mahoney. Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification. *arXiv:2007.05869*, 2020.

[158] Sicco Verwer, Azqa Nadeem, Christian Hammerschmidt, Laurens Bliek, Abdullah Al-Dujaili, and Una-May O'Reilly. The Robust Malware Detection Challenge and Greedy Random Accelerated Multi-Bit Search. In *13th ACM Workshop on Artificial Intelligence and Security*, 2020.

[159] Viet Quoc Vo, Ehsan Abbasnejad, and Damith C Ranasinghe. Query Efficient Decision Based Sparse Attacks Against Black-Box Deep Learning Models. *arXiv:2202.00091*, 2022.

[160] Dan Wang, Jiayu Lin, and Yuan-Gen Wang. Query-Efficient Adversarial Attack Based on Latin Hypercube Sampling. In *IEEE International Conference on Image Processing*. IEEE, 2022.

[161] Fangwei Wang, Yuanyuan Lu, Qingru Li, Changguang Wang, and Yonglei Bai. A Co-evolutionary Algorithm-Based Malware Adversarial Sample Generation Method. In *2022 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 1–8. IEEE, 2022.

[162] Guoqiu Wang, Xingxing Wei, and Huanqian Yan. Improving Adversarial Transferability with Spatial Momentum. *5-th Chinese Conference on Pattern Recognition and Computer Vision*, 2022.

[163] Lu Wang, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Yuan Jiang. Spanning attack: reinforce black-box attacks with unlabeled data. *Machine Learning*, 2020.

[164] Ruikui Wang, Yuanfang Guo, Ruijie Yang, and Yunhong Wang. Exploring Transferable and Robust Adversarial Perturbation Generation from the Perspective of Network Hierarchy. *arXiv preprint arXiv:2108.07033*, 2021.

[165] Xiaosen Wang and Kun He. Enhancing the Transferability of Adversarial Attacks through Variance Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[166] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the Transferability of Adversarial Attacks. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[167] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting Adversarial Transferability through Enhanced Momentum. *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[168] Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle Attack: A Query-Efficient Decision-Based Adversarial Attack. In *European Conference on Computer Vision*. Springer, 2022.

[169] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A Unified Approach to Interpreting and Boosting Adversarial Transferability. *International Conference on Learning Representations,*, 2020.

[170] Xiruo Wang and Risto Miikkulainen. MDEA: Malware Detection with Evolutionary Adversarial Learning. In *IEEE Congress on Evolutionary Computation*. IEEE, 2020.

[171] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature Importance-aware Transferable Adversarial Attacks. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[172] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. *arXiv:2002.05990*, 2020.

[173] Jing Wu, Mingyi Zhou, Shuaicheng Liu, Yipeng Liu, and Ce Zhu. Decision-based Universal Adversarial Attack. *arXiv:2009.07024*, 2020.

[174] Lei Wu, Zhanxing Zhu, Cheng Tai, et al. Understanding and Enhancing the Transferability of Adversarial Examples . *arXiv:1802.09707*, 2018.

[175] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the Transferability of Adversarial Samples via Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[176] Weibin Wu, Yuxin Su, Michael R Lyu, and Irwin King. Improving the Transferability of Adversarial Samples with Adversarial Transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[177] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating Adversarial Examples with Adversarial Networks. *arXiv:1801.02610*, 2018.

[178] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving Transferability of Adversarial Examples with Input Diversity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[179] Bo Yang, Hengwei Zhang, Zheming Li, Yuchen Zhang, Kaiyong Xu, and Jindong Wang. Adversarial example generation with AdaBelief Optimizer and Crop Invariance. *Applied Intelligence*, 2022.

[180] Jiancheng Yang, Yangzhou Jiang, Xiaoyang Huang, Bingbing Ni, and Chenglong Zhao. Learning Black-Box Attackers with Transferable Priors and Query Feedback. In *Advances in Neural Information Processing Systems*, 2020.

[181] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Boosting Transferability of Targeted Adversarial Examples via Hierarchical Generative Networks. In *European Conference on Computer Vision*. Springer, 2022.

[182] Maksym Yatsura, Jan Metzen, and Matthias Hein. Meta-Learning the Search Distribution of Black-Box Random Search Based Adversarial Attacks. *Advances in Neural Information Processing Systems*, 2021.

[183] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, 2014.

[184] Haojie Yuan, Qi Chu, Feng Zhu, Rui Zhao, Bin Liu, and Neng-Hai Yu. AutoMA: Towards Automatic Model Augmentation for Transferable Adversarial Attacks. *IEEE Transactions on Multimedia*, 2021.

[185] Junkun Yuan, Shaofang Zhou, Lanfen Lin, Feng Wang, and Jia Cui. Black-Box Adversarial Attacks Against Deep Learning Based Malware Binaries Detection with GAN. In *European Conference on Artificial Intelligence*. IOS Press, 2020.

[186] Zheng Yuan, Jie Zhang, Yunpei Jia, Chuanqi Tan, Tao Xue, and Shiguang Shan. Meta Gradient Adversarial Attack. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[187] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive Image Transformations for Transfer-based Adversarial Attack. *European Conference on Computer Vision*, 2021.

[188] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Data-Free Adversarial Perturbations for Practical Black-Box Attack. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[189] Chaoning Zhang, Gyusang Cho, Philipp Benz, Kang Zhang, Chenshuang Zhang, Chan-Hyun Youn, and In So Kweon. Early Stop And Adversarial Training Yield Better surrogate Model: Very Non-Robust Features Harm Adversarial Transferability. *OpenReview*, 2021.

[190] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving Adversarial Transferability via Neuron Attribution-Based Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[191] Jiawei Zhang, Linyi Li, Huichen Li, Xiaolu Zhang, Shuang Yang, and Bo Li. Progressive-Scale Boundary Blackbox Attack via Projective Gradient Estimation. In *International Conference on Machine Learning*. PMLR, 2021.

[192] Lan Zhang, Peng Liu, Yoonho Choi, and Ping Chen. Semantic-preserving Reinforcement Learning Attack Against Graph Neural Networks for Malware Detection. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[193] Qilong Zhang, Xiaodan Li, Yuefeng Chen, Jingkuan Song, Lianli Gao, Yuan He, and Hui Xue. Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains. *arXiv:2201.11528*, 2022.

[194] Qilong Zhang, Chaoning Zhang, Chaoqun Li, Jingkuan Song, Lianli Gao, and Heng Tao Shen. Practical No-box Adversarial Attacks with Training-free Hybrid Image Transformation. *arXiv:2203.04607*, 2022.

[195] Pu Zhao, Pin-Yu Chen, Siyue Wang, and Xue Lin. Towards Query-Efficient Black-Box Adversary with Zeroth-Order Natural Gradient Descent. In *AAAI Conference on Artificial Intelligence*, 2020.

[196] Pu Zhao, Sijia Liu, Pin-Yu Chen, Nghia Hoang, Kaidi Xu, Bhavya Kailkhura, and Xue 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.

[197] Weiwei Zhao and Zhigang Zeng. Improved black-box attack based on query and perturbation distribution. In *13th International Conference on Advanced Computational Intelligence*. IEEE, 2021.

[198] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On Success and Simplicity: A Second Look at Transferable Targeted Attacks. In *Advances in Neural Information Processing Systems*, 2021.

[199] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, and Michael Backes. Towards Good Practices in Evaluating Transfer Adversarial Attacks. *arXiv:2211.09565*, 2022.

[200] Fangtian Zhong, Xiuzhen Cheng, Dongxiao Yu, Bei Gong, Shuaiwen Song, and Jiguo Yu. MalFox: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs Against Black-Box Detectors. *arXiv:2011.01509*, 2020.

[201] Fengfan Zhou, Hefei Ling, Yuxuan Shi, Jiazhong Chen, Zongyi Li, and Qian Wang. Improving Transferability of Adversarial Examples on Face Recognition with Beneficial Perturbation Feature Augmentation. *arXiv:2210.16117*, 2022.

[202] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. DaST: Data-free Substitute Training for Adversarial Attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[203] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable Adversarial Perturbations. In *European Conference on Computer Vision*, 2018.

[204] Yao Zhu, Jiacheng Sun, and Zhenguo Li. Rethinking Adversarial Transferability from a Data Distribution Perspective. In *International Conference on Learning Representations*, 2021.

[205] Junhua Zou, Yexin Duan, Boyu Li, Wu Zhang, Yu Pan, and Zhisong Pan. Making Adversarial Examples More Transferable and Indistinguishable. In *AAAI Conference on Artificial Intelligence*, 2022.

[206] Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the Transferability of Adversarial Examples with Resized-Diverse-Inputs, Diversity-Ensemble and Region Fitting. In *European Conference on Computer Vision*. Springer, 2020.