Low-Rank and Sparse Decomposition for Low-Query Decision-Based Adversarial Attacks

Ashkan Esmaeili, *Member, IEEE*, Marzieh Edraki, Nazanin Rahnavard, *Senior Member, IEEE*, Ajmal Mian, *Senior Member, IEEE*, and Mubarak Shah, *Life Fellow, IEEE*

***Abstract*— Deep learning models are susceptible to contrived adversarial examples, even in the decision-based black-box setting where the attacker has access to the model's decisions only. Developing more efficient and practical attacks help in better understanding the limitations of deep models. It is important that attacks are crafted with limited queries to avoid suspicion. Since the required number of queries increase with dimensions, low-dimensional embeddings are attractive. This low query budget constraint is a bottleneck for learning-based and data-driven attacks which rely heavily on querying the model. We propose LSDAT, an image-agnostic non-data-driven decision-based black-box attack that exploits low-rank and sparse decomposition (LSD) of images to dramatically reduce the queries and improve fooling rates compared to existing methods. LSDAT crafts perturbations in the low-dimensional subspace formed by the sparse component of the input image and that of a target adversarial image to obtain query-efficiency. A viable perturbation is obtained by traversing the path between the input and adversarial sparse components. Theoretical analyses are provided to justify the functionality of LSDAT. Unlike other competitors (e.g., FFT), LSD works directly in the image domain to guarantee that non-ℓ_2 constraints, such as sparsity, are satisfied. LSDAT offers better control over the number of queries and is computationally efficient as it performs sparse decomposition of the input and adversarial images only once to generate all queries. Four variants of LSDAT are presented for different scenarios including a pure black-box attack where no queries are allowed. We demonstrate ℓ_0, ℓ_2 and ℓ_∞ bounded attacks with LSDAT to evince its efficiency compared to baseline attacks in diverse low-query budget scenarios. LSDAT obtains 15 to 20% improvement in fooling ResNet-50 while using far fewer queries than competing methods in a similar setting.**

***Index Terms*— Low rank and sparse decomposition, black-box attack, adversarial examples, query budget, decision based attack.**

Manuscript received 5 January 2022; revised 8 July 2022 and 12 December 2022; accepted 8 January 2023. Date of publication 12 May 2023; date of current version 19 December 2023. This work was supported in part by the National Science Foundation under Grant ECCS-1810256 and Grant CCF-1718195, and in part by the Defense Advanced Research Projects Agency under Agreement HR00112090095. The work of Ajmal Mian was supported as the recipient of an Australian Research Council Future Fellowship Award funded by the Australian Government under Project FT210100268. The associate editor coordinating the review of this manuscript and approving it for publication was Ms. Elham Tabassi. *(Corresponding author: Ashkan Esmaeili.)*

Ashkan Esmaeili and Marzieh Edraki were with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: ashkan.smiley@knights.ucf.edu).

Nazanin Rahnavard is with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL 32816 USA.

Ajmal Mian is with the Department of Computer Science, The University of Western Australia, Perth, WA 6009, Australia.

Mubarak Shah is with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816 USA.

Digital Object Identifier 10.1109/TIFS.2023.3275737

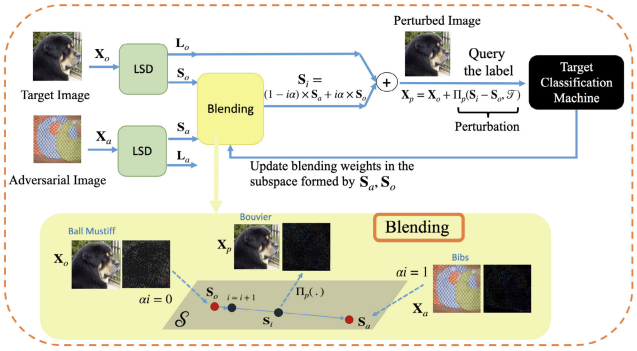


Fig. 1. Overview of LSDAT. LSD is performed to extract sparse components of the input image and some initial adversarial sample. The path between the two sparse components is then gradually traversed with a step size (α) to blend the sparse components via a weighted combination. The perturbation is projected to satisfy the imperceptibility constraint. Finally, the perturbed image is obtained by adding the blended sparse component $\mathbf{S}_i$ to the low-rank component of the input image. The target model is queried with the perturbed image. If the model is not fooled, the process repeats with increasing values of iteration index ($i = i + 1$) until the model is fooled or the query budget gets exhausted.

I. Introduction

DEEP learning has brought revolutionary change across academia, industry, and daily life. Deep neural network models can often match or surpass human performance on well-defined cognitive tasks in ideal conditions such as image understanding related tasks [49]. However, their performance sharply degrades under minor alterations to the input signals or task objective. The distinct gap in robustness between human and computational intelligence is a central challenge for the next generation of AI research. Adversarial vulnerability is a striking example of the shortcomings of deep learning. The outputs of virtually all deep learning models are extremely sensitive to small changes in the input signal. This sensitivity can be exploited to create imperceptible signals that completely disrupt network performance. The failure of computational systems to match human robustness is a crucial gap that must be bridged before AI can be deployed in security-critical contexts such as autonomous driving, medical diagnosis, malware detection, spam detection, intrusion detection, cybersecurity, video surveillance, robotics, financial services fraud detection, access control, and medical diagnosis [55]. Adversarial attacks are applicable to all deep-learning-based real-world systems such as speech recognition [11], speech-to-text conversion [13], face recognition [50], visual classification [25], malware detection [28], [31], and other general physical world cases as discussed in [5] and [36]. In this paper, we focus on black-box adversarial attacks in image understanding related tasks.

Adversarial disruptions are mainly studied to gain insights into the inner working of DNNs or to measure their robustness and resilience to adversarial disruptions, which is then useful to build and deploy a universal security system to protect AI models from imperceptible adversarial attacks. In adversarial attack, the threat is on the model (to cause misclassification) and the way to impose the threat is by creating adversarial examples. Generally, adversarial attacks can be launched in a white-box setting, where the attacker has full knowledge of complete DNN model [12], [27], [34], [40], [43], [47], as well as in a black-box setting, where access to only the model output is available [1].

The taxonomy of adversarial attack and defence scenarios in the literature progresses based on the amount of information provided to the attacker. It has been considered to be scoped down through time to less and less amount of information. The scope has been condensed from having full access to models to only having access to the data or the logits, and ultimately only the decisions of the models, i.e., the top-1 labels.

Designing an attack in this setting seems difficult as the attacker can only query the model to obtain a decision and even not a score like a logit. Many authors estimate the logits and decision boundary using several queries to the model. This made decision-based attacks in low-query budgets a recent challenge in the literature. Deep learning methods may require many queries to estimate the decision boundary with high precision. However, applying a large number of queries to a black-box model is contrary to the low-query budget assumption. Hence, in practice, when the query budget is short or in the extreme case (pure black-box setting) is zero, the attacker cannot use the learning methods, and it is inevitable to apply the model-based image-agnostic low-dimensional mappings such as Fast Fourier Transform (FFT) in order to approach the curse of dimensionality in designing the adversarial attacks. The classical signal processing model-based transforms are leveraged as they do bypass the learning procedure and further need for the queries.

To this end, we propose to employ Low-Rank and Sparse Decomposition (LSD) for decision-based black-box attack; LSDAT for brevity, which is efficient and works under a very low-query budget. The merit in using this approach is three-fold: 1- sparse perturbations are effective in fooling the classifiers, 2- they are imperceptible, 3- the sparse perturbation is image-agnostic and remains in the image original domain. Therefore, no further transformations are required as in other low-dimensional mappings (e.g., FFT) because transforms and their inverse add up to computational complexity.

The main contributions of the current work are then summarized as follows: noitemsep

- We propose an efficient decision-based black-box attack (LSDAT) that exploits the low rank and sparse decomposition of images to
  drastically reduce the number of queries.
- We introduce an online learning technique to build prior knowledge from successful attacks. Through sequential attacks, a group of prominent initial adversarial images (IAIs) are organized into 2 levels of class-specific and global dictionaries to be used as candidate IAIs for upcoming attacks. We empirically demonstrate that the top-1 entry of the global dictionary is one of the *universal* adversarial images and establish theoretical properties for such images. Exploiting the prior information significantly reduces the average queries.
- The work is buttressed with theoretical analyses which establishes why the proposed sparse perturbation is functional.
- We present variants of LSDAT matching different attack scenarios including pure black-box (zero-query) attack, hierarchical dictionary-based attack, ensemble of substitute models-based attack, and diversity-based attack.
- We provide analysis on how the computational complexity of LSDAT is less compared to other transform-based decision-based methods such as FFT-based attacks as LSDAT need not fetch the image from and to the transform domains and generates all perturbations with the one-time LSD.
- The experimental results are provided which establish the superiority of LSDAT to state-of-the-art (SOTA) in fooling rate under similar perturbation budget constraints.

II. Related Work

Adversarial attacks can be categorized in two broad groups, namely white box attacks and black-box attacks. In white box scenario, the attacker has a full access to the model architecture, parameters and data distribution. Due to non-linearity of deep learning models, many white-box adversarial attack methods rely on local information of the model such as its gradients or hessian (curvature info) in order to obtain linear or second order approximation of the model and henceforth, facilitating crafting the adversarial attack through linear or quadratic programming. This can lead to energy-efficient quasi-imperceptible perturbation computation. Deepfool [43], FGSM [27], Carlini & Wagner attack [12], Projected Gradient Attack (PGD) [40], Jacobian-based Saliency Map Attack (JSMA) [47], Elastic-Net Attack [15], and Adversarial-Bandit Attack [34] are some of the well-known attacks in this category. To this end, the authors in [43] proposed Deepfool, as an efficient instance of gradient-based methods, which works based on linear approximation of the classifier near its boundary. In [59], the authors have extended this concept and have approximated the classifier's boundary with a second-order expansion. Benefiting form a trust-region based quadratically constrained quadratic programming (QCQP), their method Trust Region Attack further ameliorates the attacker's performance via considering the local curvature information and fine-tuning the perturbation direction.

Transfer-based attacks rely on the transferability of adversarial examples among models and exploit substitute models to craft these, for example the attacks proposed in [20], [32], [39], and [46]. In this scenario, the attacker has access to the data distribution but has no information about the model. Another category is score-based attack that limits the attacker's knowledge only to the model scores such as the class probabilities or logits. The attacker tries to estimate the gradient of the model from the score through significant number of queries. References [3], [6], [16], [33], [37], [42], [44], and [54] are some of the efficient methods in this group. A query-efficient score based method is presented in [6].

Other examples of score-based attack method include Gradient Approximation QEBA [37], AdvFlow [42], AutoZOOM [54], ZOO [16], LocalSearch [44], GenAttack [3], Query-Limited Partial-Information Attack [33], and LeBA [58], which is a combination of query-based and transferability-based methods, to name a few.

Decision-based attacks are the most challenging scenario which narrows the attackers vision only to the classifier's top-1 hard label output. The first work considering decision-based attack was the Boundary Attack (BA) proposed in [8]. BA estimates the boundary and moves along it to minimize the perturbation. The Query-Limited Attack [33] leverages a Monte-Carlo approximation approach to approximate the model scores based on the label-setting only and from there on, proceeds with score-based Partial Information Attack. This can be achieved by applying several queries and averaging them to estimate the logits. In addition to inferior accuracy in estimating the logits, this approach contradicts query-efficiency as it requires even more queries to approximate the logits. Another efficient decision-based method is HopSkipJumpAttack (HJSA) [14] which is based on estimating the gradient direction using top-1 class labels. A natural evolutionary (NE) algorithm has been introduced to update the data covariance matrix after certain queries to reduce the search space from a sphere to an quadratic eclipse characterized by the covariance matrix. As the method proceeds, the empirical covariance matrix of data is updated. Assuming Gaussian prior, the algorithm becomes much faster and more efficient by reducing the search space from sphere to an eclipse characterized by the updated covariance matrix [24].

A similar approach in updating covariance matrix based on truncated Gaussian distributions is leveraged to adapt Deepfool to the decision-based GeoDA method [48]. Natural evolutionary strategies (NES) were first considered by Ilyas et al. [33] in designing query-efficient attacks. Cheng et al. [18] model the top-1 label attack as a real-valued optimization problem and use zeroth-order optimization approach to design query-efficient attack using randomized gradient-free method (RGF). Zhao et al. [62] have proposed ZO-ADMM method where they integrate the alternating direction method of multipliers (ADMM) with zeroth-order (ZO) optimization and Bayesian optimization (BO) to design a query-efficient gradient-free attack. In [19], the authors extend the optimization based approach and estimate the gradient sign at any direction instead of the gradient itself and introduce Sign-OPT which is more query-efficient compared to OPT.

Another sign-based method, SIGN-HUNTER [2] exploits a sign-based gradient approximation rather than magnitude-based to devise a binary black-box optimization. Their method does not rely on hyper-parameter tuning or dimensionality reduction. Chen, et. al. have suggested to randomly flip the signs of a small number of entries in adversarial perturbations and this way, boost the attacker's performance, specifically in defensive models compared to EA, BA, SimBA [29], SignOPT, and HSJA [17]. Reference [9] offers a bias for gradient direction based on a surrogate model. Dimension reduction based attack techniques are investigated to achieve query efficiency. Sign-OPT-FFT [19], Bayes Attack [51], SimBA-DCT [29], QEBA-S, QEBA-F, QEBA-I, and GeoDA-Subspace [48] are which are effective in ℓ_2 attacks, but in order to guarantee other imperceptibility bounds, they must fetch to the original and transformation domains consecutively, imposing computational burden on their procedure. Certain methods focus on crafting attacks which are sparse in the image original dimensions such as SparseFool [41], GreedyFool [23], Sparse attack via perturbation factorization [26] (for white-box), and Sparse-RS [22], CornerSearch [21] and GeoDA (sparse version) [48] (for black-box). The most competent method in low-query black-box attacks is the recent method Square-Attack [4]. We will consider Square-Attack which also outperforms the main rival considered in our work (Bayes Attack) in our final comparisons.

III. Proposed Method: LSDAT

The proposed LSDAT method is considered for untargeted black-box adversarial attack. Untargeted attack can be formulated as the following optimization problem:

$$\min_{\boldsymbol{\delta}} \ \|\boldsymbol{\delta}\|_p \quad s.t. \ \ \mathcal{C}(\mathbf{X}_0 + \boldsymbol{\delta}) \neq \mathcal{C}(\mathbf{X}_0), \tag{1}$$

where $\boldsymbol{\delta}$ denotes the added perturbation. The goal is to minimize the ℓ_p-norm of the perturbation such that when applied to the input image $\mathbf{X}_0$, the classifier $\mathcal{C}$ is fooled.

Since LSD is one of the main building blocks of the proposed attack, here we briefly introduce it to be self-contained. LSD is a well established optimization problem studied in classical machine learning with image and video processing applications [7], [38], [45], [63]. LSD is an image-agnostic and non-data-driven transform which assumes most images can be explained with a low-rank background plus a sparse part which is decisive in classifying the image. Mathematically, if an image is denoted by $\mathbf{X}$, LSD can be formalized as:

$$\min_{\mathbf{L},\mathbf{S}} \ \text{rank}(\mathbf{L}) + \lambda\|\mathbf{S}\|_0, \quad \text{s.t.} \ \ \mathbf{X} = \mathbf{L} + \mathbf{S}, \tag{2}$$

where $\mathbf{L}$ is the low rank and $\mathbf{S}$ is the sparse component. The regularization coefficient λ determines the sparsity level of $\mathbf{S}$. The convex surrogate functions for the rank function and the ℓ_0-norm are considered to be the trace norm $\|.\|_*$ and the ℓ_1-norm, respectively. Hence, Problem (2) can be cast as a convex optimization problem

$$\min_{\mathbf{L},\mathbf{S}} \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1, \quad \text{s.t.} \ \ \mathbf{X} = \mathbf{L} + \mathbf{S}. \tag{3}$$

There are several methods to solve the result convex programming, among which we work with Robust PCA a.k.a RPCA. Many packages can be utilized that efficiently solve (3). We use the GODEC method [63] to perform LSD.

The sparse part consists of far less (in terms of order of magnitude) non-zero pixels compared to the target image dimensions; while these reduced data are highly informative and well represent details of the image and are therefore decisive in classification [57], [60], [61]. As already stated, we present several variants of LSDAT depending on the scenario considered for the attack. For now, we present the

most primitive one which is LSDAT(R) working with only one randomly selected IAI from a pool of possible IAIs. Later, we will introduce next variants which are designed to reduce the number of required queries to reach the idea of pure black-box scenario. These include LSDAT(D), a dictionary based method, LSDST-ES (consensus-based performance on ensemble of offline models), and LSDAT-HYB (a mixture of dictionary-based and ensemble of models).

Algorithm 1 summarizes the core LSDAT approach with one IAI (LSDAT(R)). An IAI (with a ground truth label different from the target image) is randomly selected. For instance, the IAI can be chosen from the labeld validation set of the data distribution to be attacked. If no prior pool of IAIs is available such as in the pure black-box scenario, the label can be known at the cost of one query to the real-time (online) model or alternatively, annotated by another trustworthy cognition model available for the attacker in an offline mode. These different cases will shape the LSDAT variants to be scrutinized later. LSD is performed on the original and the IAI of interest, $\mathbf{X}_o$ and $\mathbf{X}_a$, respectively. Next, the sparse components of the two images, refereed to as $\mathbf{S}_o$ and $\mathbf{S}_a$, are extracted using LSD. The union of sparse components form a low-dimensional subspace $\mathcal{S}$ in the original domain. There are noticeably fewer pixels (effective dimensions) in $\mathcal{S}$ compared to the original image dimensions. This helps reducing the number of queries for crafting the perturbation (the number of queries generally scales with the dimensions justifying why low-dimensional transforms are used to craft the perturbation in a low-dimensional space). Narrowing down the attack vision to $\mathcal{S}$, we propound that for some certain IAI $\mathbf{X}_a$, Enforcing the sparsity constraint, the adversarial perturbation of interest, i.e., the sparsest vector from the original sample to the decision boundary, will be a linear (more specifically a convex) combination of $\mathbf{S}_a, \mathbf{S}_o$ that lies on the direction $\mathbf{S}_a - \mathbf{S}_o$. This will be theoretically shown in Section V.

The attack attempts to induce a new sparse pattern ($\mathbf{S}_a$) into the perturbed image $\mathbf{X}_p$ which is highly informative of $\mathbf{X}_a$ and suppresses $\mathbf{S}_o$ while traversing from $\mathbf{S}_o$ to $\mathbf{S}_a$. It is worth noting that the alteration from one sparse component to another is done via a weighted combination of both. Since the traversing is in a low-dimensional subspace, the semantic transform from $\mathbf{X}_o$ to $\mathbf{X}_a$ is carried out rapidly and within few steps (queries). In other words, small step sizes (perturbations) in the sparse domain leads to more drastic changes in the image concept. In Section V, we will show how a locally linear classifier (LLC) functions well based on a basis of low rank and sparse components and verify the efficacy of using sparse components as decisive elements in classification. Fig. 2 demonstrates this procedure.

After obtaining the perturbation on the specified direction of interest, i.e., $\mathbf{S}_a - \mathbf{S}_o$, the perturbation is projected on the ℓ_p-ball depending on the imperceptibility constraint. Projection is denoted by Π_p in Alg. 1 line 4. ℓ_p norm bounds are the prevalent perturbation constraints considered in the literature. In our work, we consider ℓ_0, ℓ_2, and ℓ_∞ norm bounds on the perturbations.

Algorithm 1 LSDAT With One Initial Adversarial Image (IAI)

Require: $(\mathbf{X}_o, r)$: target image and its class, $\mathbf{X}_a$: Initial adversarial image, MaxIter: Maximum number of iterations, α: sparse traversing rate, p: ℓ_p constraint type, $\mathcal{T}$ = Imperceptibility constraint budget $\{k, \epsilon, \sigma\}$

Output: $\mathbf{X}_p$: Perturbed image, N_Q: Number attempted queries, F_{attack}: Attack success flag

Initialization: $\mathbf{X}_p \leftarrow \mathbf{0}$, $N_Q \leftarrow 0$, $F_{attack} \leftarrow False$, $i \leftarrow 1$

1: $(\mathbf{L}_o, \mathbf{S}_o) \leftarrow \mathbf{LSD}(\mathbf{X}_o)$, $(\mathbf{L}_a, \mathbf{S}_a) \leftarrow \mathbf{LSD}(\mathbf{X}_a)$
2: **while** $i \leq$ MaxIter **do**
3: $\quad \mathbf{S}_i \leftarrow (\alpha \times i)\mathbf{S}_a + (1 - \alpha \times i)\mathbf{S}_o$
4: $\quad \mathbf{S}_i \leftarrow \mathbf{S}_o + \Pi_p(\mathbf{S}_i - \mathbf{S}_o, \mathcal{T})$
5: $\quad \mathbf{X}_i \leftarrow \mathbf{L_o} + \mathbf{S}_i$
6: $\quad c = \text{Query}(\mathbf{X_i})$
7: $\quad$ **if** $c \neq r$ **then**
8: $\quad\quad F_{attack} \leftarrow True$, $\text{N}_\text{Q} \leftarrow i$, $\mathbf{X}_p \leftarrow \mathbf{X}_i$
9: $\quad\quad$ break
10: $\quad$ **end if**
11: $\quad i = i + 1$
12: **end while**
13: **return** $\mathbf{X}_p, \text{N}_\text{Q}, \text{F}_\text{attack}$

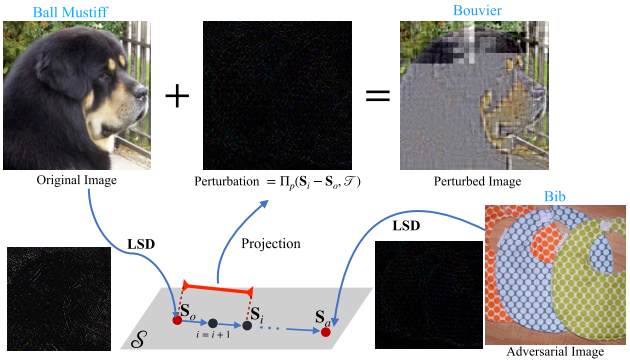


Fig. 2. Illustration of LSDAT. Perturbation lies on the path $\mathbf{S}_a - \mathbf{S}_o$ and is added in a step-wise fashion to $\mathbf{S}_o$, which is finally added to $\mathbf{X}_o$.

Finally, the perturbation $\Pi_p(\mathbf{S}_i - \mathbf{S}_o, \mathcal{T})$ is added to the image, which is defined as ℓ_p norm projection [1] of the perturbation under the imperceptibility constraint budget in Alg. 1. In Alg. 1, we use interchangeable representation by obtaining the resulted sparse part of the perturbed image denoted with $\mathbf{S}_i$. in Fig. 1, followed by adding it to the low-rank component of the target image $\mathbf{L}_o$ to form the candidate perturbed image for each query. The perturbation of interest is both effective (lies on $\mathcal{S}$) and sparse (a vertice on the ℓ_1 ball centered at $\mathbf{X}_o$).

One motivation of preferring LSD to FFT-based methods is that although transform methods like FFT-based or spatial-based approaches optimally represent geometric structure information of images, they cannot extract entire contours and edges accurately, while the LSD can extract the edges and

[1] $\Pi_2(\mathbf{V}, \epsilon) = \min\{1, \frac{\epsilon}{\|\mathbf{V}\|_2^2}\}\mathbf{V}$.
$\mathbf{V}' = \Pi_\infty(\mathbf{V}, \sigma) : v'_{ij} = \textbf{sign}\{v_{ij}\}\min\{|v_{ij}|, \sigma\} \forall (i, j)$
$\Pi_0(\mathbf{V}, k)$: Keep the largest k elements of $\mathbf{V}$ in magnitude and set the rest to zero.

salient parts of an image in an image-agnostic fashion. Side effects, such as pseudo Gibbs phenomenon and false contours are downsides of domain transform-based methods [60].

IV. Challenge of Selecting IAI

As stated previously, the perturbation direction of interest is set as $\mathbf{S}_a - \mathbf{S}_o$ for some $\mathbf{S}_a$ to maintain sparsity and small perturbation ℓ_p norm. A random choice, nevertheless, may not yield the optimal perturbation direction. Thanks to noticeable query-efficiency of LSDAT, the attacker can explore among several IAIs in a non-pure-black-box scenario. We call this set of random samples the *exploration* set, $\mathcal{E}$. If the initial adversarial image budget of an attack per sample is G, the set $\mathcal{E}$ is gathered such that $|\mathcal{E}| = G$. To launch an attack to the image $\mathbf{X}_o$, we consider one sample at a time, drawn from $\mathcal{E}$ as the IAI and apply Alg. 1 to it. Note that, if the attack is successful at any point, the rest of the exploration set will not be attempted. Thus, the number of queries is upper bounded by $(j \times MaxIter) + N_Q$, where j is the number of unsuccessful initial adversarial attempts and N_Q is the number of queries used in the successful attack.

A. Online Learning With Hierarchical Dictionaries

Black-box attack scenarios can be categorized into isolated and non-isolated ones. In non-isolated attacks, where the goal is to attack a set of images rather than a single one, the attacker can build a prior knowledge through the attacking process by learning the set of elite IAIs which are universal in fooling the previously attacked target images. These prominent samples are organized into a hierarchy of *class-specific* and *global* dictionaries. The intuition is that if one IAI functions well for multiple images of a specific class; for instance the class cat; it is also very likely to be a good initial adversarial point for the other instances of that class. If the class-specific dictionary does not have any entry or the number of entries is limited, the best IAIs are those which have successfully fooled other classes so far. All dictionaries' entries are always ranked based on their score, which is defined as the number of successful attacks for that image as an IAI up to now.

Our proposed dictionary-based attack exploits the previous good IAIs to launch a new attack with fewer queries. In a new attack on a target image $\mathbf{X}$ with label r and the IAI budget of G, first the entries of class-specific dictionary for class r are selected one after the other as the IAI. If none of them leads to an adversarial counterpart for $\mathbf{X}_o$ and the budget G is not met, the remaining initial images are picked from the global dictionary and if exhausted, from random sampling (Figure 3). If the attack is successful, we update the dictionaries accordingly, i.e. update the score for the dictionaries entries or adding a new item to them. As the attacking process continues, the dictionaries become richer to the point that top-1 entry of the global dictionary contributes significantly in successful attacks. We refer to this IAI as *universal* IAI. The properties of such images are investigated theoretically in the next section. LSDAT using dictionary is denoted as LSDAT(D) throughout the paper.

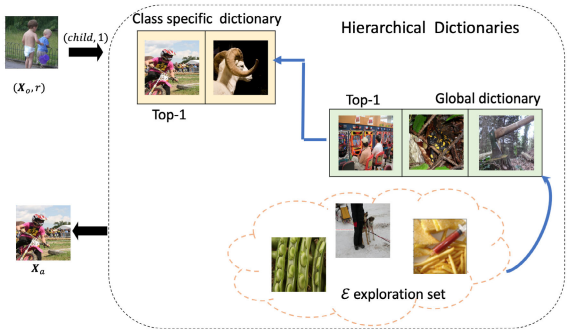


Fig. 3. Hierarchical dictionary structure for initial adversarial image provider. At each point, the image is fetched based on the class label r and the index i in the budget G.

V. Theoretical Analysis

Now, we establish the theoretical analysis for LSDAT.

We assume the original and the adversarial images are decomposed as $\mathbf{X}_o = \mathbf{L}_o + \mathbf{S}_o$, $\mathbf{X}_a = \mathbf{L}_a + \mathbf{S}_a$, where $\mathbf{L}_o, \mathbf{S}_o$ and $\mathbf{L}_a, \mathbf{S}_a$ denote the low rank and sparse components of the original and the adversarial images, respectively. The goal is to show that $\mathbf{S}_a - \mathbf{S}_o$ is a viable sparse perturbation direction centered around $\mathbf{X}_o$ that can fool the model. Before delving into the analysis, we elaborate on the geometric interpretation of LSDAT functionality, as depicted in Fig. 4. It is known that an ℓ_1-ball centered at $\mathbf{X}_o$ has sharp corners (vertices). If one gradually enlarges congruent ℓ_1-balls centered at $\mathbf{X}_o$, it is highly likely that one of them intersects the decision boundary at one of its sharp corners which is a well-known property of ℓ_1 contours.

Moreover, the ℓ_1-balls centered at $\mathbf{X_0}$ also intersect with the subspace spanned by $\mathbf{S}_o$ and $\mathbf{S}_a$ denoted as $\mathcal{S}$. The intersection can be formulated as $\sum_i s_{oi}|w_{1i}| + \sum_j s_{aj}|w_{2j}| = cte$, where s_{oi} and s_{aj} are the i and j element of $\mathbf{S}_o$ and $\mathbf{S}_a$, respectively, and w_{1i} and w_{2j} are the ℓ_1 ball parameters. A specific direction which lies on $\mathcal{S}$ and also forms a vertex for one ℓ_1-ball (due to sparsity) is $\mathbf{S}_a - \mathbf{S}_o$. The goal is to show that for some initial $\mathbf{S}_a$, traversing the path $\mathbf{S}_a - \mathbf{S}_o$ starting from $\mathbf{X}_o$ introduces a viable sparse perturbation which is highly likely to be the most aligned sparse direction with the shortest path to decision boundary (δ) compared to other vertices of the ℓ_1-ball $\|\mathbf{X} - \mathbf{X}_o\|_1 = cte$. Therefore, a perturbation lying on $\mathbf{S}_a - \mathbf{S}_o$ is both sparse and hence norm-constrained, and also likely to cross a decision boundary due to relative alignment with δ. The described concept can be visually observed in Fig. 4.

Now, we delve into mathematical analysis. First, we assume the decision boundary can be locally linearized in part of an ϵ-net covered by $\mathbf{X}_o$ and some initial $\mathbf{X}_a$ in the exploration set $\mathcal{E}$ (for larger $\|\mathcal{E}\|$, LSDAT is more likely to find such ϵ-net), where ϵ is a small value. In general, considering there are P nearest samples in an ϵ-net covering local decision boundary, a locally linear classifier (LLC) can be estimated using regression on a basis composed of low-rank and sparse components $\mathcal{B} = \{\mathbf{L}_i, \mathbf{S}_i\}_{i=1}^{P}$. The local regression weights can be used for classification (fed to a linear SVM for instance). It is shown in [61] that LLC trained on the basis formed by LSD components of nearest samples yields a favorable classifier with small generalization error. An LLC is governed by certain regression weights near each sample. This leads to

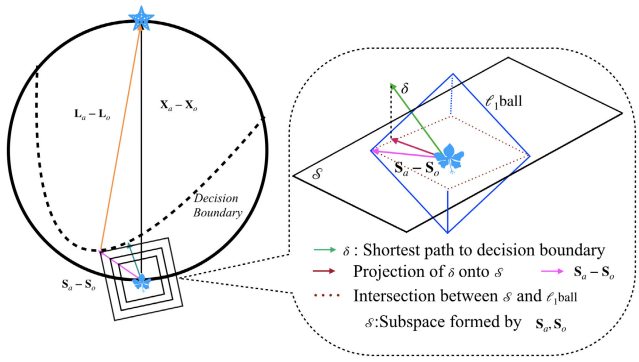


Fig. 4. Geometric illustration of LSDAT. The attempt is to show that among sparse directions, $\mathbf{S}_a-\mathbf{S}_o$ is likely to be the most aligned one with the shortest path to decision boundary δ, and therefore is likely to cross the decision boundary as well as maintaining perturbation norm efficiency.

a nonlinear classifier in general as different parameters are to be utilized for different local regions. Considering an ϵ-net covering decision boundary consisting of $\mathbf{X}_o$, the desired LLC can be obtained as follows:

$$\min_{\mathbf{c}_p} \sum_{p=1}^{P} \|\mathbf{X}_p - \mathcal{B}\mathbf{c}_p\|^2 + \lambda\|\mathbf{d}_p \odot \mathbf{c}_\mathbf{p}\|^2, \tag{4}$$

where $\mathbf{d}_p$ is defined as $\mathbf{d}_p = \exp\left(\frac{dist(\mathbf{X}_p,\mathcal{B})}{\sigma}\right)$. σ determines the decay rate for locality, and $dist(\mathbf{X}_p,\mathcal{B})$ is the distance between $\mathbf{X}_p$ and basis elements in $\mathcal{B}$. Therefore, any sample in this ϵ-net can be written using the low rank and sparse basis expansion as

$$\mathbf{X}_p = \mathcal{B}\mathbf{c}_p = \bar{\mathcal{B}}(\epsilon)\bar{\mathbf{c}}_p + \mathcal{B}(\epsilon)\mathbf{t}, \tag{5}$$

where $\mathcal{B}(\epsilon) = [\mathbf{L}_o, \mathbf{S}_o, \mathbf{L}_a, \mathbf{S}_a]$, $\bar{\mathcal{B}}(\epsilon) = \mathcal{B}\backslash\mathcal{B}(\epsilon)$, and $\bar{\mathbf{c}}_p, \mathbf{t}$ are split components of $\mathbf{c}_p$ indexing $\bar{\mathcal{B}}(\epsilon)$ and $\bar{\mathcal{B}}(\epsilon)$, respectively. The LLC prioritizes the components based on the distance from the samples. We have assumed the ϵ-net is covered by $\mathbf{X}_o$ and $\mathbf{X}_a$. Thus, they are dominant terms in LLC, and for some $\tau(\epsilon,\sigma)$ which is increasing w.r.t σ and ϵ, we have $\|\mathbf{t}\|_2^2 > \left(1-\tau^2(\epsilon,\sigma)\right)\|\mathbf{c}_p\|_2^2$. This means,

$$\|\mathbf{X}_p - \mathcal{B}(\epsilon)\mathbf{t}\| = \|\bar{\mathcal{B}}(\epsilon)\bar{\mathbf{c}}_p\| \leq \tau(\epsilon,\sigma)\|\bar{\mathcal{B}}(\epsilon)\|_{op}\|\mathbf{c}_p\| \tag{6}$$

For small σ and ϵ values, τ becomes small and the dominant components of $\mathbf{c}_p$ index $\mathcal{B}(\epsilon)$. Taking the latter into account, $\mathbf{X}_p = \mathcal{B}(\epsilon)\mathbf{t} + \mathcal{O}(\tau) \approx \mathcal{B}(\epsilon)\mathbf{t}$.

We are specifically interested in some point $\mathbf{X}_p$ (as the perturbed image) that lies in the ϵ-net and while maintaining the sparsest perturbation form $\mathbf{X}_o$, fools the model. Let $\mathbf{t} = [t_1, t_2, t_3, t_4]$. The perturbation $\mathbf{X}_p - \mathbf{X}_o$ can be written as,

$$(t_1-1)\mathbf{L}_o + (t_2-1)\mathbf{S}_o + t_3\mathbf{L}_a + t_4\mathbf{S}_a \tag{7}$$

It is desired that the perturbation 1- be sparse (or ℓ_p bounded) as much as possible, 2- be aligned with the side information of the fooling direction, i.e., $\mathbf{X}_a - \mathbf{X}_o$, as much as possible, and 3- the perturbed image $\mathbf{X}_p$ be close to $\mathbf{X_a}$ in the ϵ-net as much as possible so as to cross the boundary and fool the model. Therefore, to find the desired perturbation, the following optimization over $\mathbf{t}$ is suggested:

$$\min_{[t_1,t_2,t_3,t_4]} \underbrace{\mu\,\|(t_1-1)\mathbf{L}_o + (t_2-1)\mathbf{S}_o + t_3\mathbf{L}_a + t_4\mathbf{S}_a\|_0}_{\text{sparsity measure}} +$$

$$\underbrace{\lambda\,\theta\Big((t_1-1)\mathbf{L}_o + (t_2-1)\mathbf{S}_o + t_3\mathbf{L}_a + t_4\mathbf{S}_a, \mathbf{X}_a - \mathbf{X}_o\Big)}_{\text{alignment with the difference direction } \mathbf{L}_a+\mathbf{S}_a-\mathbf{L}_o-\mathbf{S}_o} +$$

$$\underbrace{\|t_1\mathbf{L}_o + t_2\mathbf{S}_o + (t_3-1)\mathbf{L}_a + (t_4-1)\mathbf{S}_a\|_2^2}_{\text{distance of } \mathbf{X}_p \text{ to the adversarial sample}} \tag{8}$$

where λ and μ are regularization coefficients. When μ is large enough (which is a reasonable assumption enforcing restricted perturbation norm), coefficients of $\mathbf{L}_o$ and $\mathbf{L}_a$ tend to 0 in the ℓ_1 regularized term promoting sparsity because these are largely non-sparse terms compared to $\mathbf{S}_o$ and $\mathbf{S}_a$ (similar to sparse group lasso [52]). This leads to $t_1 = 1, t_3 = 0$. Therefore, the sparsity-constrained term shrinks to $\|(t_2-1)\mathbf{S}_o + t_4\mathbf{S}_a\|_0$. Assuming orthogonality of the linear combination of $\mathbf{S}_o$ and $\mathbf{S}_a$ to $\mathbf{L}_a - \mathbf{L}_o$ on the basis $\mathcal{B}(\epsilon)$, the third term (distance of $\mathbf{X}_p$ to $\mathbf{X}_a$) can be written as $\|t_1\mathbf{L}_o + (t_3-1)\mathbf{L}_a\|_2^2 + \|t_2\mathbf{S}_o + (t_4-1)\mathbf{S}_a\|_2^2$. As stated, t_1 and t_3 values are forced by large μ.

The compromise between the controllable expression in the third term $\|t_2\mathbf{S}_o + (t_4-1)\mathbf{S}_a\|_2^2$, the sparsity regularizer $\mu\|(t_2-1)\mathbf{S}_o + t_4\mathbf{S}_a\|_0$, and the alignment term with the difference direction $\lambda\langle(t_2-1)\mathbf{S}_o + t_4\mathbf{S}_a, \mathbf{X}_a - \mathbf{X}_o\rangle$,[2] determines the weights t_2, t_4. Remembering the orthogonality assumption of sparse terms combinations to low-rank terms combinations, the alignment term can also be reduced to be expressed only based on sparse terms as $\lambda\langle(t_2-1)\mathbf{S}_o + t_4\mathbf{S}_a, \mathbf{S}_a - \mathbf{S}_o\rangle$.

The ultimate obtained programming is on variables t_2, t_4 with $\mathbf{S}_a$ and $\mathbf{S}_o$ as the determining elements. The solution therefore lies on the subspace containing the sparse parts, $\mathcal{S}$. There are two pressing reasons why the solution yields the desired direction $\mathbf{S}_a - \mathbf{S}_o$ ($t_2 = 0, t_4 = 1$). First, large λ makes the alignment term a determining one. Second, the notion of group sparsity can be applied from the beginning instead of ℓ_0 norm because $\mathbf{S}_a$ and $\mathbf{S}_o$ are already sparse and t_1 and t_3 are set to zero for large enough μ. The group sparsity term decomposed over low-rank and sparse terms leads to weighted sum of their ℓ_2 norms. Neglecting the low-rank terms, the resulted problem has only ℓ_2 norms and is hence a quadratic programming whose solution lies on the direct line between $\mathbf{S}_a$ and $\mathbf{S}_o$ depending on the compromise of the regularization coefficients. The sparsity regularizer is minimized for $t_2 = 1, t_4 = 0$, and the distance term is minimized for $t_2 = 1, t_4 = 0$. Thus, the solution lies on this direction of interest, i.e., when there is a compromise of both (large $\mu < \infty$), the solution lie somewhere on the sparse perturbation vector $(\mathbf{0}, \mathbf{S}_a - \mathbf{S}_o)$.

As mentioned, there may exists some universal samples in a dictionary of elite samples which are globally capable of fooling the model for input samples from diverse classes. Although the path $\mathbf{S}_a - \mathbf{S}_o$ has been shown to be the most aligned (best sparse approximation) sparse direction with δ, yet this alignment can vary depending on existence of the ϵ-net and the angle between δ and $\mathbf{S}_a - \mathbf{S}_o$. The angle depends on how sparse the δ is itself. Theoretically, an IAI is universally most aligned if δ is close to its sparse approximation $\mathbf{S}_a - \mathbf{S}_o$ as much as possible. As stated before,

[2] $\langle .,. \rangle$ denotes the vector inner product.

δ is expressed in the basis $\mathcal{B}(\epsilon)$ and will be the sparsest if in the representation $(t_1-1)\mathbf{L}_o+(t_2-1)\mathbf{S}_o+t_3\mathbf{L}_a+t_4\mathbf{S}_a$, we have $t_3 = 0$. Equivalently, the sample itself is almost explained by its sparse component and its low-rank component is negligible.

VI. Complexity Analysis

A general fact is that the number of queries scale with the image dimensions as each coordinate can play a role in fooling the model. To remedy the curse of dimensionality in crafting adversarial perturbation for high-dimensional data, domain transforms are applied to the target image in order to design the perturbation in a low-dimensional space. FFT-based methods such as QEBA-F [37] and Bayes-Attack [51] are the most query-efficient attack methods to the best of our knowledge. Although reducing the required queries, performing low-dimensional transforms and their inverse impose extra computational burden per each query.

The complexity of the FFT-based methods (on an $n \times n$ image) is dominated by $\mathcal{O}(N \times t \times n^2 \log_2(n))$, where N is the number of queries and t is the iterations per query for FFT and its inverse IFFT. [3] Although increasing the efficiency, such transforms come at the cost of increased query-wise complexity.

On the contrary, LSDAT merits over such methods as it only applies a one-time initial LSD for IAIs attempted from the exploration set. Next, it applies summations on sparse components in the original domain. As the sparse coding and the summation all happen in the original domain, there is no additional transform-related computational burden per query in LSDAT.

The most efficient computational complexity corresponding to RPCA is obtained by accelerated alternating projections algorithm (IRCUR) [10] which is (for $m = n$) $\mathcal{O}(Gnr^2 log_2^2(n) log_2(\frac{1}{\epsilon}))$, where r is the rank of the low-rank component, and ϵ is the accuracy of the low-rank component (appearing as the number LSD solver algorithm iterations), and G is the number of explored samples in the exploration set $\mathcal{E}$. It immediately follows that the proposed are less complex compared to transform based methods with a factor of $\frac{log(n)}{n}$ which plays an important role in high-dimensional setting. Additionally, FFT-based methods do not obtain control on non-ℓ_2 (such as ℓ_0 or ℓ_∞) perturbation constraints in the transform domain. This mandates applying extra transforms to perform clipping, thresholding, and projections back in the original domain in order for satisfying such imperceptibility constraints. Extra transforms come at the cost of more computational burden. While LSDAT directly maneuvers the image in the original domain obtaining direct control on such constraints.

VII. Experimental Set-Up and Results

In this section, we present a comprehensive set of experiments to demonstrate the efficacy of the proposed LSDAT attack in degrading the performance of well-trained image classifiers for ImageNet. Experiments are designed in diverse settings and LSDAT comes in different versions depending on the scenario of interest.

We apply LSDAT to attack two ImageNet pre-trained models,[4] namely ResNet-50 [30] and VGG16 [53] on the set $\mathcal{D}$, created by gathering correctly classified images from the ImageNet validation set. We increase the step size in LSDAT implementations for ℓ_0 and ℓ_∞ scenarios as it does not affect the corresponding constraints. Fast convergence of LSDAT with large step size allows us to expand the exploration set $\mathcal{E}$ to increase the fooling rate. From now on, we abbreviate Average Queries and Fooling Rate with AQ and FR, respectively. In the following experiments, FR is defined based on the number of mis-classified target samples divided by the number of samples in $\mathcal{D}$. The reported number of queries is averaged on all successful attack instances. We compare the performance of LSDAT based on AQ, FR, and perturbation norm with SOTA methods. We first investigate non-pure-black-box scenario where the attacker can obtain online learning and present **LSDAT(R)** and **LSDAT(D)**. In LSDAT(x), x="R" represents random IAI selection, x="D" stands for dictionary-based selection.

LSDAT with ℓ_2 constraint: For a fair comparison in ℓ_2 attack scenario, we use 1000 samples for the set $\mathcal{D}$ and the IAI budget is set to 100. For LSDAT(R), we select the IAIs randomly from validation set. Note that IAI set varies for each image to be attacked. In LSDAT(D), we exploit the prior information by first selecting IAIs from the class specific (if exists) and then the global dictionary. If the IAI budget is not met, the remaining samples are selected randomly. It is worth noting that we only add an image to the IAI set, if the current images could not lead to a successful attack so far. The comparison of performance with the SOTA methods is presented in Table I. The LSDAT attack consistently outperforms all methods with a significant drop in AQ. LSDAT(R) leads to on average 28.7% and 29.6% reduction in AQ while it improves the fooling rate by 17.38% and 31.16% for ResNet-50 and VGG models, respectively compared to Bayes attack [51]. Applying LSDAT(D), the attack further improves the FR while reducing the AQ by 47.6% for ResNet-50 and 57% for VGG compared to Bayes attack. The AQ gain is one order of magnitude compared to other methods.

VIII. Implementation of LSDAT(D)

In this section, we provide more details about the implementation of the dictionary based LSDAT. Algorithm 2 summarizes the procedure of LSDAT(D) for the query budget $G \geq 1$. While the IAI budget G is not met and the attack has not been successful, the IAI $\mathbf{X}_a$ is fetched from dictionaries (Line 2). The module *InitalAdvSmplProvider* treats the class r specific dictionary and the global dictionary as a connected array as depicted in Figure 6 and returns the sample in the index j as the adversarial sample $\mathbf{X}_a$. If j is larger than the number of samples in the connected array, the adversarial sample is randomly selected from the exploration set. The

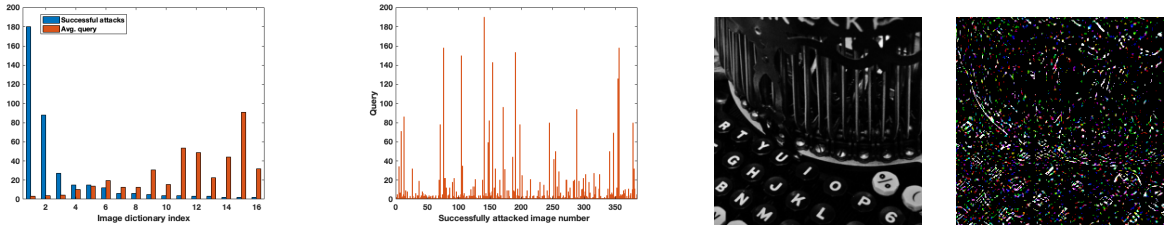
[3] In general, dimension-reduction transforms such as PCA have complexity $\mathcal{O}(mn \min\{m, n\})$. FFT is privileged over PCA due to its implementation structure.

[4] https://pytorch.org/docs/stable/torchvision/models.html

TABLE I

COMPARISON OF DIFFERENT ℓ_2 ATTACK METHODS PERFORMANCE ON IMAGENET BASED ON VARIOUS ℓ_2 BALL CONSTRAINT ϵ FOR EFFECT OF HYPER PARAMETERS. FR AND AQ STAND FOR FOOLING RATE AND AVERAGE QUERY RESPECTIVELY. IN LSDAT(X), X="R" REPRESENTS RANDOM SAMPLES, X="D" STANDS FOR DICTIONARY BASE. BEST PERFORMANCES ARE IN BOLD

Method	ResNet-50						VGG-16-BN					
	$\epsilon = 5$		$\epsilon = 10$		$\epsilon = 20$		$\epsilon = 5$		$\epsilon = 10$		$\epsilon = 20$	
	FR	AQ	FR	AQ	FR	AQ	FR	AQ	FR	AQ	FR	AQ
BA [8]	8.52	666.5	15.39	577.9	26.97	538.1	11.23	626.3	21.27	547.6	39.37	503.2
OPT Attack [18]	7.64	777.4	15.84	737.2	32.53	757.9	11.09	736.6	21.79	658.9	43.86	718.7
HJSA [14]	6.99	904.3	14.76	887.1	28.37	876.8	10.30	893.2	21.53	898.2	40.82	892.6
Sign-OPT [19]	7.46	777.4	15.84	737.1	32.53	757.9	19.81	841.1	35.8	843.7	60.63	857.7
Bayes Attack [51]	20.10	64.2	37.15	64.1	66.67	54.97	24.04	69.8	43.46	76.5	71.99	48.9
LSDAT(R)	*23.40*	*53.9*	*47.6*	*41.5*	*75.20*	*35.6*	*30.20*	*58.8*	*55.6*	*43.8*	*81.00*	*33.9*
LSDAT(D)	**25.40**	**35.2**	**47.6**	**39.4**	**76.80**	**21.50**	**32.80**	**36.4**	**56.80**	**32.9**	**82.40**	**15.2**



(a) (b) (c) (d)

Fig. 5. From left to right,(a) Successful attacks (blue bars) and average query distribution of global dictionary samples with success rate> 1. The orange bar represents the average query per successful attack. (b) The number of queries for each successful attack. (c) The top-1 IAI of the global dictionary. (d) The sparse component of the top-1 IAI which is scaled for the sake of visibility.

TABLE II

COMPARISON OF PERFORMANCE OF LSDAT L_∞ ATTACK FOR $\sigma = 0.05$ WITH SOTA METHODS

Method	ResNet-50		VGG16-bn	
	FR	AQ	FR	AQ
OPT Attack [18]	5.73	246.3	7.53	251.2
Sign-OPT [19]	10.31	660.4	15.85	666.6
Bayes Attack [51]	67.48	45.9	**78.47**	**33.7**
LSDAT(R)	**70.00**	31.3	76.20	43.2
LSDAT(D)	69.40	**29.4**	74.80	37.3

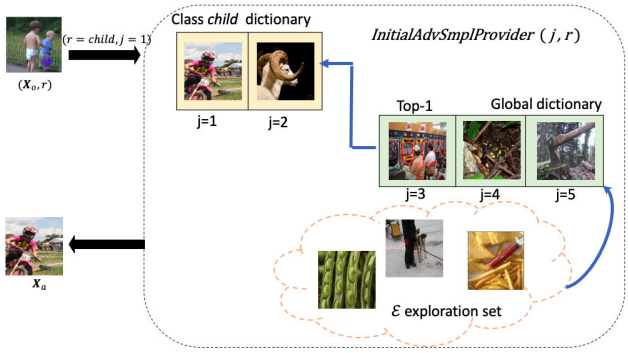


Fig. 6. Schematic of IAI provider module for the class child. The class specific dictionary and the global dictionary are linked together as a connected array. The adversarial sample is fetched from this array based on the index j.

attack is launched based on the adversarial sample $\mathbf{X}_a$ and if it is successful, the dictionaries get updated (Line 6). This algorithm returns the perturbed image $\mathbf{X}_p$, the total number of queries for the attack Q_t and the attack success flag A_f.

1) Universal Adversarial Sparse Image: We also analyze the effectiveness of dictionary in reducing the AQ and finding the universal sparse IAI. To this end, we apply LSDAT(D) attack with ℓ_2 constraint of $\epsilon = 20$ on a set with $|\mathcal{D}| = 500$.

Algorithm 2 LSDAT(D) With IAI Budget $G \geq 1$

Require: $(\mathbf{X}_o, r)$: target image and its class, G: IAIs budget, MaxIter: Maximum number of iterations, α: sparse traversing rate, p: ℓ_p constraint type, $\mathcal{T}$ = Imperceptibility constraint budget $\{k, \epsilon, \sigma\}$

Output: $\mathbf{X}_p$: Perturbed image, Q_t: Total number of queries, A_f: Attack success flag

Initialization: $Q_t \leftarrow 0$, $A_f \leftarrow False$, $j \leftarrow 1$

1: **while** $j \leq$ G and A_f *is* $False$ **do**
2: $\mathbf{X}_a \leftarrow$ InitialAdvSmplProvider(j, r)
3: $\mathbf{X}_p, N_Q, A_f \leftarrow LSDAT(\mathbf{X}_o, r, \mathbf{X}_a, MaxIter, \alpha, p, \mathcal{T})$
4: **if** $A_f == True$ **then**
5: $Q_t = (j-1) \times MaxIter + N_Q$
6: UpdateDictionary$(\mathbf{X}_a, r)$
7: break
8: **end if**
9: $j = j + 1$
10: **end while**
11: **return** $\mathbf{X}_p, Q_t, A_f$

The initial sample budget is 100 with *MaxIter=2* per sample. This setting leads to 384 successful attacks. Figure 5-(a) shows the distribution of dictionary samples that bring about at least 2 successful attacks in blue bars along with the AQ per attack for each image in orange bars. Manifestly, the top-1 IAI in the global dictionary gives rise to 46.8% of successful attacks with as low as 3.27 query per attack and the top-5 samples are responsible for 86.7% of successes with average of 7.11 queries per attack. These findings support the existence of a universal IAI with a dominant sparse component whose difference to the input sparse component is highly likely to

TABLE III

COMPARISON OF LSDAT ℓ_0 ATTACK PERFORMANCE TO RESNET-50 MODEL UNDER VARIOUS PERTURBATION RATES (P%) WITH GEODA. IN LSDAT(X), X="R" REPRESENTS RANDOM IAIS, X="D" STANDS FOR DICTIONARY BASE

P%	Method					
	GeoDA [48]		LSDAT(R)		LSDAT(D)	
	FR	AQ	FR	AQ	FR	AQ
4.29	88.44	500	85.20	12.6	**90.00**	**8.3**
3.05	82.30	500	80.20	15.3	**83.40**	**8.6**
2.36	75.20	500	76.80	15.3	**80.10**	**10.0**
1.00	47.00	500	60.60	17.6	**64.00**	**12.2**
0.50	30.00	500	49.80	24.2	**51.20**	**17.2**

TABLE IV

COMPARISON OF LSDAT ℓ_0 ATTACK PERFORMANCE TO AN ADVERSARIALLY ROBUST RESNET-50 MODEL UNDER VARIOUS PERTURBATION RATES (P%) WITH GEODA. IN LSDAT(X), X="R" REPRESENTS RANDOM IAIS, X="D" STANDS FOR DICTIONARY BASE

Perturbation %	Method	FR	AQ
4.29	GeoDA	71.3	500
	LSDAT(R)	73	19.7
	LSDAT(D)	73	9.8
3.05	GeoDA	60.1	500
	LSDAT(R)	65	21.3
	LSDAT(D)	62.2	13.1
2.36	GeoDA	54.7	500
	LSDAT(R)	60	24.0
	LSDAT(D)	58	10.8
1.00	GeoDA	36.8	500
	LSDAT(R)	44	27.2
	LSDAT(D)	43	18.7
0.50	GeoDA	22.6	500
	LSDAT(R)	30.0	32.2
	LSDAT(D)	26.0	21.1

TABLE V

COMPARISON OF LSDAT ℓ_0 ATTACK PERFORMANCE TO RESNET-50 MODEL UNDER VARIOUS PERTURBATION RATES WITH GEODA

Method	Perturbation %	FR	AQ
GeoDA	4.29	**88.44**	500
LSDAT	4.29	87.00	**90.4**
GeoDA	3.05	82.30	500
LSDAT	3.05	**85.20**	**82.6**
GeoDA	2.36	**75.20**	500
LSDAT	2.36	73.00	**112.1**
GeoDA	1.00	47.00	500
LSDAT	1.00	**54.20**	**99.7**
GeoDA	0.50	30.00	500
LSDAT	0.50	**45.00**	**128.9**

align with the shortest path from $\mathbf{X}_o$ to the decision boundary (δ). The top-1 IAI and its sparse component are illustrated in the last 2 images of Figure 5 respectively. Clearly, the sparse component contains most of the details including keys while the background(texture) is black. This property of universal IAI is also discussed in section V.

2) LSDAT With ℓ_∞ or ℓ_0 Constraint: In case of ℓ_∞ attack, the attack setting is the same as ℓ_2 constrained attack. Table II summarizes the performance comparison of LSDAT with SOTA methods when the ℓ_∞ perturbation bound is $\sigma = 0.05$. The proposed attack, consistently outperforms all methods for ResNet-50 architecture while it achieves similar results as the runner-up method for the VGG16-bn architecture. We believe the lower performance on VGG16-bn roots in the ability of the model in extracting richer features by considering both local and global spatial information, compared to ResNet which makes the attack more difficult.

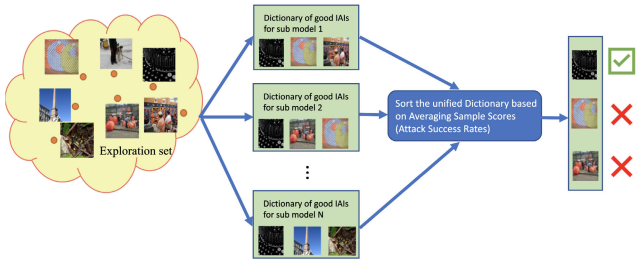
Finally, we compare LSDAT attack in ℓ_0 scenario with GeoDA [48] in Table XI. GeoDA achieves the best FR with limited query budget compared to other sparse attacks such as Sparse-RS [21], [22]. Also, Bayes Attack [51] is not the first choice to apply for ℓ_0 constraint as it suits ℓ_2 and the sparsity level is less controllable in frequency domain due to FFT transformation mandating computational burden and further queries. To have a fair comparison with GeoDA, the set $\mathcal{D}$ contains 500 samples with the IAI budget $G = 100$. Both, LSDAT(R) and LSDAT(D) significantly outperform GeoDA and improve the AQ by at least one order of magnitude. Also, the superiority of LSDAT is clear in highly imperceptible ℓ_0 attacks when only $0.5\% - 1\%$ of coordinates are perturbed. FR is improved up to 21.2% and 17% by perturbing only 0.5% and 1% coordinates respectively, setting the SOTA performance for the imperceptible ℓ_0 attacks.

3) Attacking Adversarialy Robust Models: We also evaluate the effectiveness of LSDAT against adversarially robust models. To this end, we consider the method proposed by [56] for fast adversarial training that leads to a robust ResNet-50 classifier on ImageNet with 43% robust accuracy on PGD attacks. The result of comparison of LSDAT with GeoDA with various perturbation rate for ℓ_0 constraint attacks are reported in Table VIII. While LSDAT(D) achieves higher FR with significantly lower AQ compared to GeoDA, we noticed that LSDAT(R) slightly outperforms LSDAT(D) in terms of FR. This phenomena is expected as the adversarial training changes the shape of decision boundary and makes the dictionary entries with small score less reliable as IAIs. This necessitates finding a balance factor between exploration set and exploiting dictionary. We postpone this study to our future works.

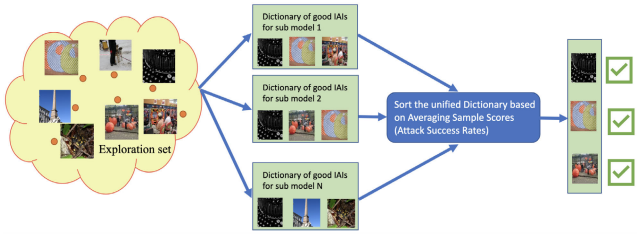
IX. ADAPTING LSDAT FOR PURE BLACK-BOX (ZERO-QUERY) ATTACK

The most challenging type of black-box attacks is known as pure black-box attack in which *only one query* is allowed to launch an attack. Before presenting our customized LSDAT version for this attack scenario, we report the performance for the primitive LSDAT version, LSDAT(R) under this attack scenario. LSDAT(R) achieves FR of 25.8% and 33.6% on ResNet-50 and VGG, respectively, for ℓ_2 constraint of $\epsilon = 20$. With *only* 1% perturbation on ℓ_0 constraint attacks, the FR=24.4% for ResNet and FR=21% for VGG can be obtained. Finally on ℓ_∞ attack with constraint of $\sigma = 0.05$ the FR is 26% and 25.4% on ResNet and VGG, respectively. Note that other decision-based black-box attacks are not applicable in this threat models as they demand more than one query to estimate the decision boundary. For instance, GeoDA [48] requires at least 10 queries to obtain average ℓ_2 distance of 39.4 which is as twice as LSDAT with a single query.

It is worth noting that although LSDAT(D) and LSDAT(R) can draw samples from the set $\mathcal{D}$ and spend more than one query, however, LSDAT can function only with one IAI suiting it for the pure black-box scenario as presented in Alg. 1.



(a) LSDAT-ES procedure



(b) LSDAT-HYB procedure

Fig. 7. Elaborating how diversity increases fooling success rate for LSDAT-HYB vs. LSDAT-ES.

TABLE VI

FOOLING RATE (FR) FOR TRANSFERRING PERTURBATIONS FROM SUBSTITUTE MODELS TO THE ZERO-QUERY BLACK-BOX MODEL ATTACK WITH LSDAT-ES

Substitute Models → target model (LSDAT-ES)	SSIM	FR
ResNet-34	0.8	41.4%
ResNet-34+ResNet-18	0.8	**48.1**%
ResNet-34+ResNet-18+DenseNet-121	0.8	47.3%
ResNet-34+ResNet-18+DenseNet-121+MobileNet-V3	0.8	44.4%
ResNet-34	0.9	24.1%
ResNet-34+ResNet-18	0.9	**27.5**%
ResNet-34+ResNet-18+DenseNet-121	0.9	27.1%
ResNet-34+ResNet-18+DenseNet-121+MobileNet-V3	0.9	26.4%

TABLE VII

FOOLING RATE (FR) FOR TRANSFERRING PERTURBATIONS FROM SUBSTITUTE MODELS TO THE ZERO-QUERY BLACK-BOX MODEL ATTACK WITH LSDAT-HYB

Substitute Models → target model	SSIM	FR
ResNet-34	0.9	26.3%
ResNet-34+ResNet-18	0.9	**28.3**%
ResNet-34+ResNet-18+DenseNet-121	0.9	24.2%
ResNet-34+ResNet-18+DenseNet-121+MobileNet-V3	0.9	24.2%

In this section, we present a customized LSDAT approach designed for the ultimately pure black-box attack scenario where the attacker cannot perform but one real-time query to fool the classifier, i.e., the pure black-box attack (a.k.a zero-query attack). To this end, we assume that the attacker can use offline substitute models to obtain universally transferable perturbations that can be used without any knowledge on the model or the target data prior distribution.

A. Zero-Query Attacks Using Ensemble of Substitute Models

In zero-query attacks, queries are not allowed at the attack time for designing the perturbations. Therefore, the proposed hierarchical dictionary-based attack is not acceptable for the zero-query scenario. To this end, we propose to exploit diverse substitute models and employ LSDAT to find fooling rates for all exploration set IAI samples. Afterwards, the exploration set samples are sorted based on their average fooling performance. Finally, the top-scorer sample is used for the real-time attack. We call this method **LSDAT-ES**. The explained procedure can be shown in Fig. 7a.

The key concept is using different DNN architectures whose functions vary in practice. It is expected that as much as the substitute model architecture approaches the real model, the proposed attack success rate on the black-box model improves. The substitute models are desired to mimic and imitate the performance of the black-box model as much as possible.

However, this is a constraining assumption as the real black-box model can take various architectural designs. In order to enhance the probability of acceptable performance in generalizing the attack on substitute models to the real model, we propose to utilize several popular architectures to broaden the scope of attacked models such that the attacker performance is universally successful regardless of the model.

With that being said, we consider several architectures, specifically ResNet-18, ResNet-34, MobileNet-V3 and DenseNet-121. These four substitute models are pretrained on the scrapped data. We train these models on a different data than ImageNet (target samples) so as to measure the generalizability of the perturbations to a different dataset.

The target model considered in our setting is the ResNet50 pretrained on ImageNet data. We have conducted the following set of experiments.

B. Experimental Setups for Zero-Query Attack With Offline Ensemble of Substitute Models

In this section, we explain the experimental setup for the proposed method with ensemble of substitute models. We randomly pick IAIs to form an exploration set. Next, for each substitute model, we form a dictionary of sorted IAIs based on their scores in fooling that model. After averaging the IAI performances across all substitute models, the top-1 scorer is considered for the final zero-query attack on the real model. In this experiment, we have incrementally added the substitute models to evaluate the ensemble model's performance as follows in four stages:

1- ResNet-18 (substitute A)

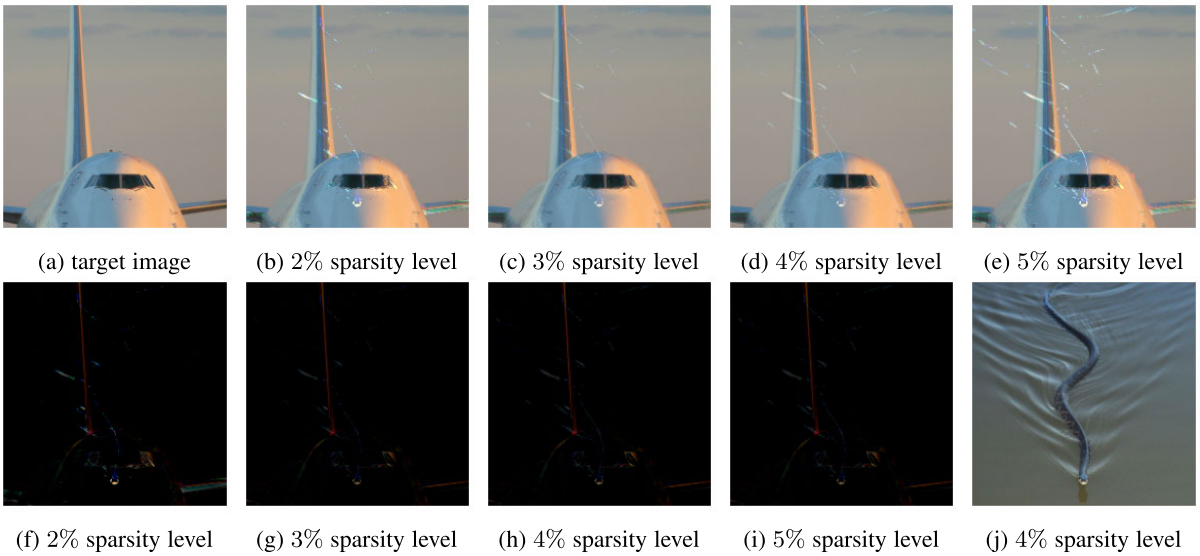
2- ResNet-34 + ResNet-18 (substitute A+B)

3- ResNet-34 +ResNet-18 + DenseNet-121 (substitute A+B+C)

4- ResNet-34+ResNet-18+DenseNet121+MobileNetV3 (substitute A+B+C+D).

We also briefly explain the experimental setup in forming the ensemble dictionary. We consider 500 samples to be attacked and the exploration set contains 200 randomly selected samples from all classes. The IAI samples are drawn uniformly to score the samples fairly (Each IAI has approximately the same number of trials). Tables VI summarizes the simulation results. We report the fooling rate for given structural similarity index measures (SSIM) as the imperceptibility measure.

We conclude that considering an ensemble of substitute models can improve the fooling rate. However, including network architectures which do not resemble the target model may endorse sparse patterns which are not effective on the



(a) target image (b) 2% sparsity level (c) 3% sparsity level (d) 4% sparsity level (e) 5% sparsity level

(f) 2% sparsity level (g) 3% sparsity level (h) 4% sparsity level (i) 5% sparsity level (j) 4% sparsity level

Fig. 8. An example of ℓ_0 attack using LSDAT with different sparsity levels.

target model and may downgrade the overall performance. On the other hand, keeping the substitute models whose expressible features are close to those of the target model will better generalize the top-1 scorer as the effective attacker selected from the dictionary.

C. Customizing the Attack to the Target Samples

Top-1 IAI, as stated in Section IX-A, may fail to work on a sequence of similar target samples that are to be fooled consecutively. This can lead to significant drop in the fooling rate. To avoid this, we may insert the notion of diversity as follows. Instead of disposing all but the top-1 scorer of the substitute model dictionaries, we keep a stack of the elite ones. We can look upon this stack to optimize the sample optimally acting on the target to be fooled. In the zero-query attack, a search through the stack is not doable although the stack size is too small. Instead, to maintain efficiency and introducing diversity, we randomly choose the IAI from the stack. The difference between the stack and the exploration set is that the stack contains all elite samples which are all likely to perform well. The merit of using stack over one single universal sample is that stack samples may not all fail on a series of similar samples which are difficult to be fooled with the top-1 scorer. In other words, stack IAIs union is a comprehensive set capable of fooling a wide range of target images. We set the stack size to 8. The procedure is illustrated in Fig. 7b. We call this method the LSDAT-HYB. the simulation results for the similar setting as in Section IX-B.

Finally, we conclude that the LSDAT-HYB method outperforms LSDAT-ES thanks to inserting the diversity as elaborated earlier which is also clear according to the results in Tables VI & VII.

X. Ablation Study

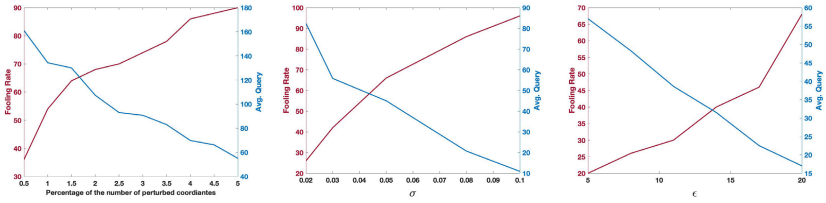
In this subsection, we investigate the effect of certain hyper-parameters on the performance of the LSDAT. Figure 9a shows the AQ and FR versus the percentage of coordinates which are allowed to be modified for ResNet50 model. The expected behavior is that the FR increases and AQ decreases with the percentage of the coordinates allowed to be perturbed. We observe that for 5% ℓ_0 perturbation budget, less than 60 AQ suffice to achieve 90% FR. Having only 0.5% of coordinates to perturb, more than 35% FR is obtained with less than 200 AQ. A visualized example of ℓ_0 attack can be found in Fig. 8.

In Fig. 9b, LSDAT achieves more than 25% FR with around 80 AQ for a harsh ℓ_∞ perturbation limit of $\sigma = 0.02$. Relaxing the ℓ_∞ perturbation constraint, we obtain 96% FR with less than 11 AQ. The effect of hyper parameter ϵ in case of ℓ_2 constrained attacks is presented in 9c as well. Hyper-parameters are varied in a common valid range to be comparable with other works fairly. We evaluate the effect of sparse traversing rate α and IAI budget G on the performance of LSDAT. We also study the effect of transferring dictionaries between attacks.

A. Sparse Traversing Rate α

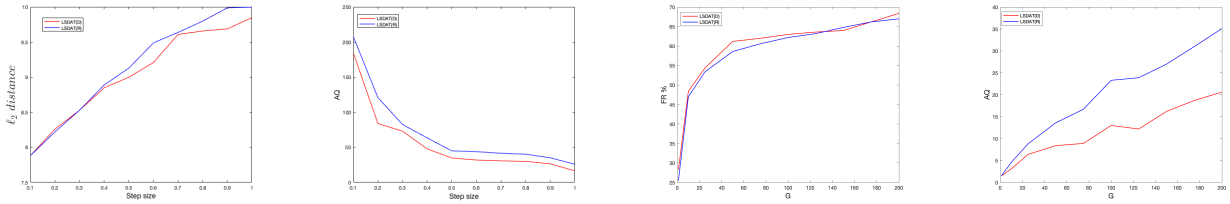
As we mentioned in the theoretical analysis (section V), the sparse traversing rate (step size) mostly affects the LSDAT with ℓ_2 constraint. To this end we study the effect of changing α in the range of [0.1, 1] on the ℓ_2 distance and AQ of LSDAT on the pre-trained ResNet-50 model on ImageNet dataset. In this set of experiments, the $|\mathcal{D}| = 500$, $G = 100$. Note that, with the small step size α, the $MaxIter$ parameter should grow to guarantee the full traverse from S_o to S_a if need be.

To this end, we set the $MaxIter$ as $\frac{2}{\alpha}$ so it increases accordingly for the small step sizes. The FR of LSDAT(R) and LSDAT(D) are close to each other with the average of 48.36% and 48.6% respectively. Figure 10a and Figure 10b plot the effect of step size on ℓ_2 distance and AQ for LSDAT(R) (blue curve) and LSDAT(D) (red curve). As it can be seen, increasing the step size negatively affect the imperceptibility of the attack by increasing the ℓ_2 distance while it improves the AQ. So there is always a trade off between these two



(a) ℓ_0 (b) ℓ_∞ (c) ℓ_2

Fig. 9. Effect of hyper parameters on the performance of LSDAT. a) Fooling rate and Avg. query versus percentage of the number perturbed coordinates (pixels). b) Fooling rate and Avg. query versus σ. c) Fooling rate and Avg. query versus ϵ.



(a) (b) (c) (d)

Fig. 10. (a) Effect of the step size on AQ of LSDAT with ℓ_2 constraint $\epsilon = 10$ on ImageNet with IAI budget G=100. (b) Effect of the step size on AQ of LSDAT with ℓ_2 constraint $\epsilon = 10$ on ImageNet with IAI budget G=100. (c) Effect of the initial adversarial budget G, on FR of LSDAT with ℓ_0 constraint $p = 1\%$ for ImageNet. (d) Effect of the initial adversarial budget G, on AQ of LSDAT with ℓ_0 constraint $p = 1\%$ for ImageNet.

factors that should be adjusted based on the threat model. For the minimum AQ, the maximum step size is required and for the best imperceptibility the minimum step size is favorable.

B. Transferring Dictionaries

We also study the transferibility of the prior knowledge between model architectures. To this end we use dictionaries that are generated by LSDAT(D) with ℓ_2 constraint $\epsilon = 10$ for attacking VGG-16-BN model tranied on ImageNet to launch a new set of attacks on Renset-50 model. The attacks parameters are as follow: $|\mathcal{D}| = 1000$, $G = 100$, step size $\alpha = 50$ and $MaxIter = 4$. We refer to this type of attack as LSDAT(TD) as TD stands for transferred dictionaries. Table IX reports the result of this study. Using the transferred dictionaries by LSDAT(TD) significantly reduces the AQ compared to LSDAT(D) as expected. Since the high score samples of the transferred dictionary can lead to successful attacks with few number of queries. However, the FR also drops since low score entries of the dictionaries might not be a better initial adversarial point for the ResNet-50 model than exploring a new random sample.

C. IAI Budget G

We also analyze the effect of IAI budget G on the FR and AQ. To this end, we used LSDAT with ℓ_0 constraint of 1% perturbation to attack ResNet-50 model trained on ImageNet dataset. In this set of experiments, $|\mathcal{D}| = 500$, step size $\alpha = 1$ and $MaxIter = 2$. The IAI budget G varies in $\{1, 10, 25, 50, 75, 100, 125, 150, 175, 200\}$. Note that increasing G leads to increasing the maximum query budget for an attack with a fixed $MaxIter$ parameter. For instance with $G = 1$ and $MaxIter = 2$ the maximum query budget is 2 per attack while with $G = 200$, we have the maximum query budget of 400. Figures 10c and 10d show the effect of G on the FR and AQ respectively. Manifestly, the fooling rate (FR) increases for larger G but it doesn't follow a linear trend of improvement. Increasing G from 1 to 100 improves the FR by 36% while further increasing to 200 IAIs only improves FR around 5%. However, the AQ has a linear relation with G. As the G grows, the effect of using hierarchical dictionaries become more evident. Comparison of the gap between AQ of LSDAT(R) and LSDAT(D) shows that exploiting prior knowledge from the dictionaries can significantly reduces the AQ even when the number of maximum query budget is large.

XI. Concluding Remarks

A query-efficient decision-based adversarial attack (LSDAT) is introduced based on low-rank and sparse decomposition. The method is suitable for very limited query budgets and is of low complexity compared to SOTA. LSDAT is also effective in fooling rate dominating the SOTA in performance as verified through diverse set of experiments. LSDAT finds a sparse perturbation which is likely to be aligned with the sparse approximation of the shortest path from input sample to the decision boundary. We show the path lies on the path connecting original and some adversarial sparse components. Theoretical analyses buttresses LSDAT performance in fooling. As few pixels entail the image information in the sparse component and the number of queries is relevant to the effective dimension of image to be fooled, the proposed method acts as a query-efficient attack. Moreover, LSDAT offers better control over imperceptibility constraints in the original domain and less complexity compared to SOTA as it does not apply consecutive transforms and their inverse. Unlike other dimension reduction techniques which craft the perturbation in the transform domain and therefore, lose control on image ℓ_p properties while crafting, LSDAT finds the perturbation in the original domain to have control on imperceptibility constraints and has complexity and faster convergence rate. Several

TABLE VIII

COMPARISON OF LSDAT ℓ_0 ATTACK PERFORMANCE TO AN ADVERSARIALY ROBUST RESNET-50 MODEL UNDER VARIOUS PERTURBATION RATES (P%) WITH GEODA. IN LSDAT(X), X="R" REPRESENTS RANDOM INITIAL ADVERSARIAL SAMPLES, X="D" STANDS FOR DICTIONARY BASE

P%	Method					
	GeoDA [48]		LSDAT(R)		LSDAT(D)	
	FR	AQ	FR	AQ	FR	AQ
4.29	71.30	500	73.00	19.71	**73.00**	**9.8**
3.05	60.10	500	**65.00**	20.3	62.20	**13.1**
2.36	54.70	500	**60.00**	24.0	58.10	**10.8**
1.00	36.80	500	**44.00**	27.2	43.00	**18.7**
0.50	22.60	500	**30.00**	32.2	26.20	**21.1**

variants of LSDAT including LSDAT(D), LSDAT(R), LSDAT-ES, and LSDAT-HYB are presented depending on the attack scenario and IAI budget. The series is designed to ideally reach the zero-query attack, aka pure black-box attack.Experiments on the well-known ImageNet dataset shows query efficiency and fooling rate superiority of LSDAT compared to SOTA.

APPENDIX A

A. Attacking Adversarially Robust Models

We also evaluate the effectiveness of LSDAT against adversarially robust models. To this end, we consider the method proposed by [56] for fast adversarial training that leads to a robust ResNet-50 classifier on ImageNet with 43% robust accuracy on PGD attacks. The result of comparison of LSDAT with GeoDA with various perturbation rate for ℓ_0 constraint attacks are reported in Table VIII. While LSDAT(x) achieve higher FR with significantly lower AQ compared to GeoDA, we noticed that LSDAT(R) slightly outperforms LSDAT(D) in terms of FR. This phenomena is expected as the adversarial training changes the shape of decision boundary and makes the dictionary entries with small score less reliable as initial adversarial images. This necessitates finding a balance factor between exploration set and exploiting dictionary. We postpone this study to our future works.

B. Pure Black-Box Attack

The most challenging type of black-box attacks is known as pure black box attack in which *only one query* is allowed to launch an attack. We evaluated the performance of the LSDAT in this scenario. LSDAT(R) achieves FR of 25.8% and 33.6% on ResNet-50 and VGG respectively for ℓ_2 constraint of $\epsilon = 20$. With *only* 1% perturbation on ℓ_0 constraint attacks, the FR=24.4% for ResNet and FR=21% for VGG can be obtained. Finally on ℓ_∞ attack with constraint of $\sigma = 0.05$ the FR is 26% and 25.4% on ResNet and VGG, respectively. Note that other decision-based black box attacks are not applicable in this threat models as they demand more than one query to estimate the decision boundary. For instance, GeoDA [48] requires at least 10 queries to obtain average ℓ_2 distance of 39.4 which is as twice as LSDAT with a single query.

APPENDIX B

In this part, we discuss the comparison between the most recent method for low-query black-box adversarial attack, *Square Attack* [4], and LSDAT. We use the reported performance for Square Attack in [4]. To provide a fair comparison, we report the same settings for both methods. As the ℓ_2 and ℓ_∞ cases are included in [4] on ResNet-50 and VGG-16-BN, we report these two scenarios accordingly. Table X summarizes the performance of LSDAT(D) vs. Square attack for the mentioned scenarios. The ℓ_2 constraint is set to $\epsilon = 5$ and the ℓ_∞ constraint is set to $\sigma = 0.05$ for both methods, respectively. Moreover, for fair compariosn, the non-isolated attack (one IAI trial) version of LSDAT, i.e., LSDAT(R) is considered. As can be observed the LSDAT(R) outperforms Square-Attack for extremely low queries. We report the values for Square-Attack based on Figure 4 in [4]. As we observe, in the ℓ_2 attack scenario, LSDAT(R) achieves higher fooling rate for smaller or equal number of queries. In the ℓ_∞ case, LSDAT outperforms in ResNet-50 and fails to compete Square-ATtack on VGG-16-BN. The reason underlies in how susceptile each model is to a sparse perturbation. In other words, it depends on how the extracted features rely on the sparse patterns (usually local patterns) or not.

TABLE IX

COMPARISON OF LSDAT(X) WITH ℓ_2 CONSTRAINT OF $\epsilon = 10$ ON THE PRETRAINED IMAGENET RESNET-50 MODEL. IN LSDAT(X), "R" STANDS FOR THE RANDOM EXPLORATION SET, "D" STANDS FOR DICTIONARY BASED ATTACK AND "TD" REPRESENTS THE TRANSFERRED DICTIONARY SCENARIO. HERE THE DICTIONARIES ARE TRANSFERRED FROM VGG-16-BN MODEL TO RESNET-50 MODEL

Method	FR%	AQ
LSDAT(R)	47.6	41.5
LSDAT(D)	47.6	39.4
LSDAT(TD)	36.8	17.3

APPENDIX C

Experiments on CIFAR-10 dataset In this section, we include the results of the proposed LSDAT on the CIFAR-10 [35] dataset. The CIFAR-10 dataset consists of 50,000 training and 10,000 test samples which are considered as the validation set in our experiments. We have considered 300 randomly selected samples form the CIFAR-10 validation set for perturbation.

We compare the fooling rate achieved with average queries (AQ) used to yield successful attack as well as the perturbation constraint. One can find that LSDAT dominates the fooling rate of HJSA with only 2-3 AQ. This is thanks to the fact that the method does not rely on estimating the decision boundary and the gradient, rather relies on the ad-hoc image-agnostic LSD decomposition and swaps the sparse patterns in one single step to make queries. As discussed in the theoreticl analysis, the mentioned perturbation direction is informative of the shortest path and is the most sparse perturbation aligned with it, therefore the direction is likely to fool. Given that LSDAT(R) and LSDAT(D) are query-efficient, the query-bidget allows them to exploring for finding an initial adversarial sample whose sparse component in functional in fooling the model, the FR of LSDAT outweighs that of HSJA significantly with far less queries. It is notable that the average ℓ_2 distance

TABLE X

COMPARISON OF SQUARE ATTACK AND LSDAT(R) FOR ℓ_2 AND ℓ_∞ SCENARIOS. THE PERTURBED DATA IS IMAGENET. PERFORMANCE ON RESNET-50 AND VGG-16-BN ARE REPORTED. FR AND AQ STAND FOR FOOLING RATE AND AVERAGE QUERY, RESPECTIVELY. BEST PERFORMANCES ARE IN BOLD

Method	ResNet-50 $\epsilon = 0.1$		VGG-16-BN $\epsilon = 0.1$		ResNet-50 $\sigma = 0.05$		VGG-16-BN $\sigma = 0.05$	
	FR	AQ	FR	AQ	FR	AQ	FR	AQ
Square Attack	< 30	50	< 30	36.4	< 70.00	35	$>$ **80**	40
LSDAT(R)	**32.8**	36.	**32.8**	36.4	**70.00**	31.3	76.20	43.2

TABLE XI

COMPARISON OF LSDAT ℓ_0 ATTACK PERFORMANCE TO RESNET-50 MODEL UNDER VARIOUS PERTURBATION RATES (P%) WITH GEODA. IN LSDAT(X), X="R" REPRESENTS RANDOM INITIAL ADVERSARIAL SAMPLES, X="D" STANDS FOR DICTIONARY BASE

P%	Method					
	GeoDA [48]		LSDAT(R)		LSDAT(D)	
	FR	AQ	FR	AQ	FR	AQ
4.29	88.44	500	85.20	12.6	**90.00**	**8.3**
3.05	82.30	500	80.20	15.3	**83.40**	**8.6**
2.36	75.20	500	76.80	15.3	**80.10**	**10.0**
1.00	47.00	500	60.60	17.6	**64.00**	**12.2**
0.50	30.00	500	49.80	24.2	**51.20**	**17.2**

TABLE XII

FOOLING RATE (FR) AND AVERAGE QUERY (AQ) FOR ℓ_∞ ATTACK ON CIFAR-10 DATASET

method	FR	AQ	ℓ_∞ distance	type
LSDAT(R)	80.1%	12	0.03	untargeted
LSDAT(D)	**82.6**%	**10.7**	0.03	untargeted
HSJA	60%	1K	0.03	untargeted
Sign-OPT	31.87%	679.39	0.03	untargeted
Bayes Attack	70.38%	75.88	0.03	untargeted
Sign-OPT	3.50%	937.65	0.03	targeted
Bayes Attack	48.93%	149.15	0.03	targeted
LSDAT(R)	56.71%	23.9	0.03	targeted
LSDAT(R)	81.8%	11.9	0.05	untargeted
LSDAT(D)	**96.9**%	**11.9**	0.05	untargeted
HSJA	84%	1K	0.05	untargeted

for LSDAT is upper bounded with the distance introduced by HSJA. Table XII draws a comparison between LSDAT(R) and HJSA, Sign-OPT, and Bayes Attack in untargeted and targeted settings. Although our method is designed to work best for untargeted attacks, we also include some evaluations on targeted setting for integrity of the reports on CIFAR-10 dataset. For untargeted attack, it can be observed that the proposed method outperforms HSJA with 22% and Bayes Attack with 12% in FR, with considerably fewer queries. It is worth noting that the dictionary of universal samples favors in this scenario as LSDAT(D) outperforms LSDAT(R). Regarding the targeted setting, we can definitely observe significant drop in the fooling rate compared to the untargeted setting which is due to the design of our method, i.e., the sparsest shortest path to decision boundary may not align with some pre-specified decision boundary (targeted attack). Yet, we observe that LSDAT(R) achieves an FR comparable to that of Bayes Attack (slightly outperforming) showing the FR is also noticeable within limited number of queries for targeted setting. In general, it is worth noting that as there are more classes, the untargeted attack is more probable to yield higher fooling rate using LSDAT, because the probability that some initial adversarial sample exists in the ϵ-net increases, and hence, it is more probable that the candidate initial sample which falls close to the decision boundary in an ϵ-net exists. In addition, more classes introduce more adjacent decision boundaries for an input image to be perturbed and therefore the probability of crossing one of them using the introduced is enhanced.

REFERENCES

[1] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[2] A. Al-Dujaili and U.-M. O'Reilly, "Sign bits are all you need for black-box attacks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–25.

[3] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "GenAttack: Practical black-box attacks with gradient-free optimization," in *Proc. Genet. Evol. Comput. Conf.*, Jul. 2019, pp. 1111–1119.

[4] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 484–501.

[5] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 284–293.

[6] A. N. Bhagoji, W. He, B. Li, and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, 2018, pp. 158–174.

[7] T. Bouwmans, N. S. Aybat, and E.-H. Zahzah, *Handbook of Robust Low-Rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. Boca Raton, FL, USA: CRC Press, 2016.

[8] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," 2017, *arXiv:1712.04248*.

[9] T. Brunner, F. Diehl, M. T. Le, and A. Knoll, "Guessing smart: Biased sampling for efficient black-box adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4958–4966.

[10] H. Cai, K. Hamm, L. Huang, J. Li, and T. Wang, "Rapid robust principal component analysis: CUR accelerated inexact low rank estimation," *IEEE Signal Process. Lett.*, vol. 28, pp. 116–120, 2021.

[11] N. Carlini et al., "Hidden voice commands," in *Proc. 25th USENIX Secur. Symp. (USENIX Security)*, 2016, pp. 513–530.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[13] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 1–7.

[14] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1277–1294.

[15] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," 2017, *arXiv:1709.04114*.

[16] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 15–26.

[17] W. Chen, Z. Zhang, X. Hu, and B. Wu, "Boosting decision-based black-box adversarial attacks with random sign flip," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 276–293.

[18] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," 2018, *arXiv:1807.04457*.
[19] M. Cheng, S. Singh, P. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh, "Sign-OPT: A query-efficient hard-label adversarial attack," 2019, *arXiv:1909.10773*.
[20] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 10934–10944.
[21] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein, "Sparse-RS: A versatile framework for query-efficient sparse black-box adversarial attacks," 2020, *arXiv:2006.12834*.
[22] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4724–4732.
[23] X. Dong et al., "GreedyFool: Distortion-aware sparse adversarial attack," 2020, *arXiv:2010.13773*.
[24] Y. Dong et al., "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7714–7722.
[25] K. Eykholt et al., "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.
[26] Y. Fan et al., "Sparse adversarial attack via perturbation factorization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 35–50.
[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[28] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," 2016, *arXiv:1606.04435*.
[29] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," 2019, *arXiv:1905.07121*.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] W. Hu and Y. Tan, "Black-box attacks against RNN based malware detection algorithms," 2017, *arXiv:1705.08131*.
[32] Z. Huang and T. Zhang, "Black-box adversarial attack with transferable model-based embedding," 2019, *arXiv:1911.07140*.
[33] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," 2018, *arXiv:1804.08598*.
[34] A. Ilyas, L. Engstrom, and A. Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," 2018, *arXiv:1807.07978*.
[35] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18268744
[36] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.
[37] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "QEBA: Query-efficient boundary-based blackbox attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1221–1230.
[38] X. Liu, G. Zhao, J. Yao, and C. Qi, "Background subtraction based on low-rank and structured sparse decomposition," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2502–2514, Aug. 2015.
[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," 2016, *arXiv:1611.02770*.
[40] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
[41] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "SparseFool: A few pixels make a big difference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9087–9096.
[42] H. M. Dolatabadi, S. Erfani, and C. Leckie, "AdvFlow: Inconspicuous black-box adversarial attacks using normalizing flows," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–14.
[43] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.
[44] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1310–1318.
[45] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components," *Magn. Reson. Med.*, vol. 73, no. 3, pp. 1125–1136, Mar. 2015.
[46] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.
[47] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroSP)*, Mar. 2016, pp. 372–387.
[48] A. Rahmati, S.-M. Moosavi-Dezfooli, P. Frossard, and H. Dai, "GeoDA: A geometric framework for black-box adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8446–8455.
[49] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017.
[50] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," 2018, *arXiv:1801.00349*.
[51] S. N. Shukla, A. K. Sahu, D. Willmott, and J. Z. Kolter, "Simple and efficient hard label black-box adversarial attacks in low query budget regimes," 2020, *arXiv:2007.07210*.
[52] S. Noah, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Stat.*, vol. 22, no. 2, pp. 231–245, May 2013.
[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[54] C.-C. Tu et al., "AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 742–749.
[55] Y. Vorobeychik and M. Kantarcioglu, "Adversarial machine learning," *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 12, no. 3, pp. 1–169, 2018.
[56] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," 2020, *arXiv:2001.03994*.
[57] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[58] J. Yang, Y. Jiang, X. Huang, B. Ni, and C. Zhao, "Learning black-box attackers with transferable priors and query feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12288–12299.
[59] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. W. Mahoney, "Trust region based adversarial attack on neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11350–11359.
[60] B. Zhang, X. Lu, H. Pei, Y. Liu, W. Zhou, and D. Jiao, "Multi-focus image fusion based on sparse decomposition and background detection," *Digit. Signal Process.*, vol. 58, pp. 50–63, Nov. 2016.
[61] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, and S. Ma, "Image classification by non-negative sparse coding, low-rank and sparse decomposition," in *Proc. CVPR*, Jun. 2011, pp. 1673–1680.
[62] P. Zhao et al., "On the design of black-box adversarial examples by leveraging gradient-free optimization and operator splitting method," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 121–130.
[63] T. Zhou and D. Tao, "GoDec: Randomized low-rank & sparse matrix decomposition in noisy case," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 33–40.