

Towards Few-Call Model Stealing via Active Self-Paced Knowledge Distillation and Diffusion-Based Image Generation

Vlad Hondru

Department of Computer Science, University of Bucharest
Bucharest, Romania
vlad.hondru25@gmail.com

Radu Tudor Ionescu*

Department of Computer Science, University of Bucharest
Bucharest, Romania
raducu.ionescu@gmail.com

ABSTRACT

Diffusion models showcased strong capabilities in image synthesis, being used in many computer vision tasks with great success. To this end, we propose to explore a new use case, namely to copy black-box classification models without having access to the original training data, the architecture, and the weights of the model, *i.e.* the model is only exposed through an inference API. More specifically, we can only observe the (soft or hard) labels for some image samples passed as input to the model. Furthermore, we consider an additional constraint limiting the number of model calls, mostly focusing our research on few-call model stealing. In order to solve the model extraction task given the applied restrictions, we propose the following framework. As training data, we create a synthetic data set (called proxy data set) by leveraging the ability of diffusion models to generate realistic and diverse images. Given a maximum number of allowed API calls, we pass the respective number of samples through the black-box model to collect labels. Finally, we distill the knowledge of the black-box teacher (attacked model) into a student model (copy of the attacked model), harnessing both labeled and unlabeled data generated by the diffusion model. We employ a novel active self-paced learning framework to make the most of the proxy data during distillation. Our empirical results on two data sets confirm the superiority of our framework over two state-of-the-art methods in the few-call model extraction scenario.

CCS CONCEPTS

• **Security and privacy** → **Software reverse engineering**; • **Computing methodologies** → *Computer vision problems*; *Active learning settings*; *Semi-supervised learning settings*.

KEYWORDS

Model stealing, knowledge distillation, diffusion models, few-shot learning, active learning, self-paced learning.

ACM Reference Format:

Vlad Hondru and Radu Tudor Ionescu. 2023. Towards Few-Call Model Stealing via Active Self-Paced Knowledge Distillation and Diffusion-Based Image Generation. In *Proceedings of Arxiv (Preprint)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Image classification is one of the most studied topics in computer vision. The task has been extensively investigated [12, 16, 24], and

*Corresponding author.

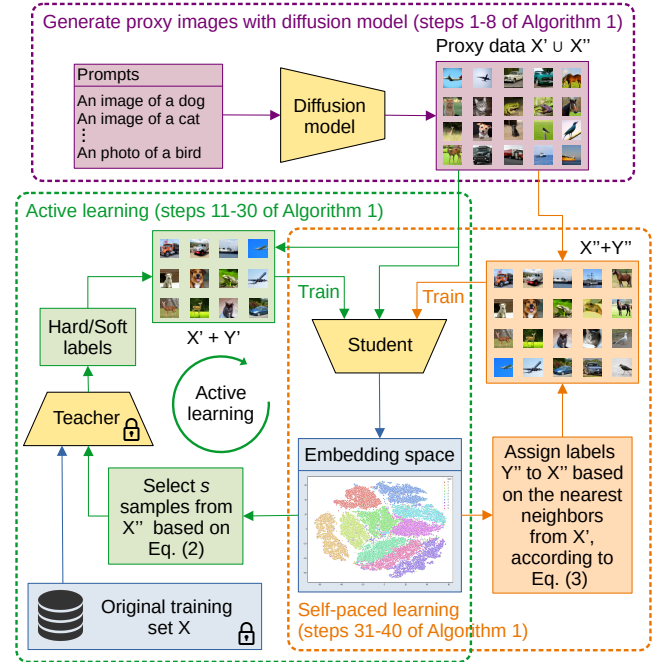


Figure 1: The proposed pipeline for model stealing starts by generating proxy images using a diffusion model. Then, proxy images are gradually annotated by the black-box teacher model and further used to train the student model via an active learning scheme. At the same time, the remaining proxy images are pseudo-labeled via a nearest neighbor scheme that operates in the latent space of the student. The pseudo-labeled images are also used to optimize the student via a self-paced learning scheme. Best viewed in color.

as a result, there is a vast amount of open-source models that can easily be accessed, even by non-technical people. However, these are usually trained on popular data sets (*e.g.* CIFAR-10 [23] or ImageNet [41]), being constrained to only predict specific object classes. If other classes are of interest, one might use a subscription-based model made available by some company, usually via a paid API. Another solution would be to train a task-specific model instead of using an already available one, the drawback being the need for a large quantity of annotated data and hardware resources.

With the recent AI hype, more and more individuals and businesses are eager to try or even implement AI-based solutions. In this context, enterprises ranging from small companies to large corporations have deployed deep learning models and made them

publicly available. In most cases, such models are accessible as Machine Learning as a Service (MLaaS), although, on a few occasions, the models and the weights are open-sourced. The most popular example is ChatGPT [34], which was made available by OpenAI. As far as the cost is concerned, a subscription-based payment scheme is often employed. Nevertheless, the facile access to the APIs results in many potential risks. One potential vulnerability is that the model’s functionality can be copied [7, 9, 32, 35, 37, 43, 48, 49], which infringes the intellectual property of the owners. As a result, exploring different methods on how to replicate black-box models will increase awareness on the existing risks, and will foster the inclusion of prevention mechanisms in the incipient development phases to counter model extraction attacks.

Given the aforementioned context, we present a pipeline that extracts the functionality of a black-box classification model (named teacher) into a locally created copy model (called student) via knowledge distillation [1, 7, 27, 29, 53] and self-paced active learning, as shown in Figure 1. Our method is deemed to be applicable in challenging real-world scenarios, where information about the data set and the training methodology (architecture, optimizer, hyperparameters, weights or other involved techniques) used to train the attacked model is completely concealed from the attacker. More precisely, our method is only able to observe the soft or hard class labels for a limited number of given input samples.

As illustrated in Figure 1, the first stage of our pipeline is to employ a diffusion model to generate a proxy data set with the samples that belong to the classes of interest. Diffusion models [10] are a type of probabilistic generative models that gained a lot of traction given their ability to outperform generative adversarial networks (GANs) [11]. These models were successfully applied to a wide range of tasks [10], ranging from unconditional image generation [17, 31, 46, 47], inpainting [26, 30] and text-to-image generation [3, 40, 42] to image segmentation [2, 4] and medical imaging [50]. The wide adoption of diffusion models is determined by their capability of generating realistic, qualitative and diverse images. To the best of our knowledge, we are the first to employ diffusion models to generate proxy data for model stealing attacks.

The next stage is to gather soft labels from the teacher model for a subset of the generated images. The size of this subset is constrained by the limited number of allowed API calls. We also consider the scenario when the black-box model returns only hard labels, showing that our pipeline is robust to the type of available labels. We propose a clustering-based approach to determine which samples are more relevant to be passed through the teacher. This is achieved by creating clusters in the latent space of the student model for each class, and computing a sampling probability for every data point, according to the distance to its corresponding cluster centroid. Then, we train our student model in a supervised setting on the labeled samples, until convergence. Finally, we introduce a self-paced learning method in which we assign pseudo-labels to the samples that were not inferred by the API due to the limited number of calls. We further train the student on a joint data set containing data samples with labels from the teacher, as well as pseudo-labeled examples. To the best of our knowledge, we are the first to study few-call model stealing.

We conduct experiments on two image data sets, CIFAR-10 [23] and Food-101 [6], considering various convolutional architectures

for the teacher and student models. As a result of our experiments, we conclude that our pipeline generally outperforms competing methods [7, 35] by significant margins, regardless of the number of API calls. We further confirm the applicability of our method in real scenarios by showing similar efficiency, irrespective of the architecture or the type of output given by the black-box model.

We summarize our contributions on replicating black-box classification models below:

- We harness diffusion models to create synthetic proxy data sets consisting of relevant samples for model stealing attacks.
- We propose a novel strategy on how to actively choose the samples for which to collect labels from the attacked model, obtaining improved results in the few-call model stealing scenario.
- We introduce a novel strategy that assigns pseudo-labels to the left-over samples and uses them to further boost the performance of the student via self-paced learning.

2 RELATED WORK

The model stealing research directions can be divided into different categories from multiple perspectives. For instance, related studies can be divided based on their main goal into attacking [7, 9, 35, 37, 43] methods and defense methods [19, 21, 25, 52, 54]. Another organization is given by the trade-off between accuracy [7, 9, 44] and number of API calls [8, 38, 45, 48]. Moreover, some studies [32, 48, 49] are aimed at retrieving exact information about the attacked model, *e.g.* its architecture or its hyperparameters, while others [7, 18, 35, 44] are aimed at mimicking its behavior. Two important categories in which the model stealing methods can be divided are given by the data used for training. Some methods [9, 36, 37] use real data, just as the attacked model, while others [7, 20, 28, 35, 43, 51] assume the training data is not accessible, resorting to artificially generating proxy data. We refer the readers to the survey of Oliynyk *et al.* [33], who presented a comprehensive taxonomy comprising multiple model stealing methods, which are described in great detail. We next concentrate on closely related studies that are replicating black-box models by launching attacks, while taking into consideration the balance between accuracy and number of model queries.

In one of the earliest works in this area, Tramér *et al.* [48] showed how to extract the capability of a black-box model, but instead of just copying the functionality, their goal was to approximate the parameters. In order to carry out such a strict task, they assumed some prior insight about the model type and training data. They advocate using the exact classes (hard labels) as the output of the attacked models to greatly improve the prevention of stealing attacks. In contrast, we demonstrate similar performance levels irrespective of the output type (soft or hard), while preserving the black-box nature of the teacher.

A stepping stone in model stealing research was the paper from Papernot *et al.* [37], which presented a method that had a similar setting as ours, but with a different objective: instead of trying to fully replicate the black-box model functionality with high accuracy, they are approximating the decision boundary. While also leveraging synthetic data obtained by augmenting some part of the original data set (thus weakening one of our assumptions), their

aim is to only launch adversarial attacks. Related efforts have been made by Biggio *et al.* [5] and Goodfellow *et al.* [15], but with even weaker constraints on knowledge about the data and the teacher model.

Bărbălu *et al.* [7] developed a framework, called Black-Box Ripper, that generates samples using GANs and then optimizes the samples with an evolutionary algorithm until the images become relevant, *i.e.* produce a high response from the teacher model. Although the presented results showed better performance than alternative approaches, the authors assumed a relaxed setting, in which an unbounded number of API calls is permitted. For a fair comparison with Black-Box Ripper, we consider the same number of API calls for both Black-Box Ripper and our framework.

Similar to Black-Box Ripper [7], Sanyal *et al.* [43] and Xie *et al.* [51] leveraged GANs to create synthetic samples which are subsequently used in launching stealing attacks. Nevertheless, the latter authors simultaneously trained the generative and the discriminative models, thus continuously improving the quality of the artificial data. Sanyal *et al.* [43] demonstrated their method for a larger number of classes (100), as well as using only hard labels. Xie *et al.* [51] chose a different approach by implementing an active learning strategy for the classes to be sampled by the GANs.

With the same objective as our work, Orekandy *et al.* [35] introduced Knockoff Nets, an approach to replicate a deep learning model made available as MLaaS, focusing at the same time on being mindful with respect to the number of queries. They utilized a large-scale proxy data set, namely ImageNet [41], and, in order to make as few API calls as possible, they employed a reinforcement learning strategy that trains a policy to choose the more relevant samples.

Different from the aforementioned related works, we do not require any additional data to obtain the proxy data samples. Moreover, we take a step further in regards to the number of permitted API calls, and not only try to minimize them, but rather have a fixed low number of queries. To the best of our knowledge, we are the first to propose a few-call model stealing framework that is applicable in all respects to a real model theft scenario.

3 METHOD

We begin by presenting the studied task and continue by introducing our method for replicating black-box models, while describing our novel components and how to integrate them in the proposed framework.

Problem statement. As stated in previous works [35], the model stealing task is very similar to knowledge distillation, *i.e.* in both cases, the task is to infuse the functionality of a teacher model into a student model. However, the goal of knowledge distillation is to produce a compressed model with comparable accuracy, which is different from the goal of black-box model stealing. In the context of model stealing, we assume no knowledge about the training data, the architecture and the weights of the teacher. Moreover, the student architecture is not required to be less complex. Nevertheless, in a similar manner, we refer to the black-box model as the teacher, and the copy model as the student.

Black-box models are usually available as MLaaS. In a real scenario, service providers do not disclose any information about

the model. The training data, the architecture of the model, its weights, gradients or hyperparameters, and other related details are unknown to the MLaaS users. Furthermore, for each query, the providers only supply the output of the model, either as soft labels (class probabilities) or hard labels. We consider an even more strict scenario where the number of queries is limited due to the following consideration: the model stealing attack might get detected due to the high number of API calls. Moreover, even if the attack remains undetected, the costs might rise to unjustifiable levels after a certain number of API calls.

Formally, we can formulate the problem statement using the following objective:

$$\min_{\theta_S} \|T(X, \theta_T) - S(X' \cup X'', \theta_S)\|, \text{ subject to } |X'| \leq n, \quad (1)$$

where T is the black-box teacher model, S is the student model in which we distill the knowledge, θ_T and θ_S are their corresponding weights, X is the original data set, while X' and X'' represent the two parts for the synthetic (proxy) data set, namely the part labeled by T and the part with pseudo-labels. The aim is to optimize the student weights such that the difference between the outputs of the two models is negligible, subject to making at most n passes through the teacher T , *i.e.* n represents the number of API calls. Following previous work [1, 7, 35], instead of using models accessible via APIs, we train the teacher ourselves, prior to launching the attack. During the attack, we use the teacher in a black-box regime, thus preserving all the constraints mentioned above. We hereby attest that no information about the teacher is leaked while training the student.

Overview. Our framework comprises three stages, as illustrated in Figure 1. In the first stage, proxy images are generated by a diffusion model. In the second stage, a number of proxy images are passed to the teacher and the resulting labels are used to train the student via knowledge distillation. In the third stage, the left proxy samples are pseudo-labeled via a nearest neighbors scheme. The second and third stages are repeated in a loop until $|X'| = n$, thus generating a novel active self-paced knowledge distillation (ASPKD) framework. The three stages are formally integrated into Algorithm 1. We next describe the individual stages, referring to the corresponding steps of the algorithm along the way.

Data generation. The first challenge to overcome in order to address black-box model stealing is to procure training data. One solution is to leverage generative models to create synthetic data. While previous works [7, 43] used GANs [14] to generate proxy data samples, we resort to the use of diffusion models. Since we need to generate instances of specific object classes, we opt for text-conditional diffusion models. To thoroughly validate our method, we employ two different diffusion models: Stable Diffusion [40] and GLIDE [30]. Stable Diffusion is based on a latent diffusion model, where the diffusion process is carried out in the latent space of a U-Net auto-encoder. The U-Net integrates a cross-attention mechanism to condition the image synthesis on text. GLIDE is a diffusion model that can alternate between two guidance methods, a classifier-free method and CLIP-based method [39]. In our approach, we select the former option. We use the publicly released GLIDE model, which was trained on a heavily filtered data set.

Algorithm 1: Active Self-Paced Knowledge Distillation (ASPKD)

Input: T - the black-box teacher model, S - the student model, G - the text-conditional diffusion model, m - the number of proxy samples, C - the set of classes, n - the maximum number of teacher calls, s - the number of teacher calls per iteration ($s \leq n$), k - the number of neighbors for pseudo-labeling.

Output: θ_S - the trained weights of the student (copy) model.

- 1 $\mathcal{T} \leftarrow \{\text{"An image of a \%s"}, \text{"An photo of a \%s"}\}; \triangleleft$ initialize the set of prompt templates
- 2 $X' \leftarrow \emptyset, Y' \leftarrow \emptyset; \triangleleft$ initialize the first proxy training subset and the corresponding set of labels given by the teacher
- 3 $X'' \leftarrow \emptyset, Y'' \leftarrow \emptyset; \triangleleft$ initialize the second proxy training subset and the corresponding set of pseudo-labels
- 4 **foreach** $i \in \{1, 2, \dots, m\}$ **do**
- 5 $c \sim \mathcal{U}(C); \triangleleft$ randomly sample a class label from a uniform distribution over the set of classes
- 6 $t \sim \mathcal{U}(\mathcal{T}); \triangleleft$ randomly sample a prompt template
- 7 $x'_i \leftarrow G(t \% \text{str}(c)); \triangleleft$ replace placeholder in template t with class name c and generate an image for the given text prompt
- 8 $X'' \leftarrow X'' \cup \{x'_i\}, Y'' \leftarrow Y'' \cup \{c\}; \triangleleft$ add the generated image and the corresponding class label to the pseudo-labeled proxy subset
- 9 $\theta_S \sim \mathcal{N}\left(0, \frac{2}{d_{in}+d_{out}}\right); \triangleleft$ initialize weights of student using Xavier initialization [13]
- 10 **repeat**
- 11 $Z'' \leftarrow \emptyset; \triangleleft$ initialize the set of latent vectors
- 12 $\mu_c \leftarrow \mathbf{0}_d, v_c \leftarrow 0, \forall c \in C; \triangleleft$ initialize the class centroids, where d is the latent space dimension, and the number of samples per class
- 13 **foreach** $i \in \{1, 2, \dots, |X''|\}$ **do**
- 14 $z''_i \leftarrow \tilde{S}(x''_i, \theta_S); \triangleleft$ obtain the latent vector for sample x''_i
- 15 $Z'' \leftarrow Z'' \cup \{z''_i\}; \triangleleft$ add latent vector to the set Z''
- 16 $\mu_{y''_i} \leftarrow \mu_{y''_i} + z''_i, v_{y''_i} \leftarrow v_{y''_i} + 1; \triangleleft$ add the latent vector to the centroid of class y''_i and increase the count of samples belonging to class y''_i
- 17 $\mu_c \leftarrow \frac{\mu_c}{v_c}, \forall c \in C; \triangleleft$ compute the centroids for all classes
- 18 $\mathcal{P} \leftarrow \emptyset; \triangleleft$ initialize the set of probabilities for inclusion in X'
- 19 **foreach** $i \in \{1, 2, \dots, |X''|\}$ **do**
- 20 $c \sim \mathcal{U}(C); \triangleleft$ randomly sample a class label from a uniform distribution over the set of classes
- 21 $p_i \leftarrow \exp\left(-\frac{\Delta(\tilde{S}(x''_i), \mu_c)}{2 \cdot \sigma^2}\right); \triangleleft$ apply Eq. (2)
- 22 $\mathcal{P} \leftarrow \mathcal{P} \cup \{p_i\}; \triangleleft$ add probability to \mathcal{P}
- 23 **foreach** $i \in \{1, 2, \dots, \min\{s, n - |X'|\}\}$ **do**
- 24 $x'_i \sim \mathcal{P}(X''); \triangleleft$ sample image using the probability distribution given through \mathcal{P}
- 25 $y'_i \leftarrow T(x'_i); \triangleleft$ obtain the target label from the teacher
- 26 $X' \leftarrow X' \cup \{x'_i\}, Y' \leftarrow Y' \cup \{y'_i\}; \triangleleft$ add image and teacher label to the first proxy subset
- 27 $X'' \leftarrow X'' - \{x'_i\}, Y'' \leftarrow Y'' - \{y'_i\}; \triangleleft$ remove image and label from the pseudo-labeled proxy subset
- 28 **repeat**
- 29 **foreach** $i \in \{1, 2, \dots, |X'|\}$ **do**
- 30 $\theta_S \leftarrow \theta_S - \eta \cdot \nabla \mathcal{L}(x'_i, y'_i, \theta_S); \triangleleft$ train the student on the current version of X' with labels Y' , using the learning rate η
- 31 **until convergence;**
- 32 **foreach** $x''_i \in X''$ **do**
- 33 $D \leftarrow \mathbf{0}_{|X'|}; \triangleleft$ initialize the vector of distances with zeros
- 34 **foreach** $x'_j \in X'$ **do**
- 35 $d_j \leftarrow \Delta_{cos}(\tilde{S}(x''_i), \tilde{S}(x'_j)); \triangleleft$ apply Eq. (3) and store distance to component d_j of D
- 36 $*, I \leftarrow \text{sort}(D); \triangleleft$ sort the distances (in ascending order) and return the sorted indexes
- 37 $y''_i \leftarrow \sum_{j=1}^k (1 - d_{I_j}) \cdot y'_j; \triangleleft$ assign the label to x''_i based on a weighted average of the nearest k neighbors from X'
- 38 **repeat**
- 39 **foreach** $(x, y) \in (X' \cup X'', Y' \cup Y'')$ **do**
- 40 $\theta_S \leftarrow \theta_S - \eta \cdot \nabla \mathcal{L}(x, y, \theta_S); \triangleleft$ train the student on the current version of $X' \cup X''$ with labels $Y' \cup Y''$, using the learning rate η
- 41 **until convergence;**
- 42 **until** $|X'| = n;$

As far as the prompts are concerned, for each class, we consider two alternative prompt templates, namely “An image of a $\{class\}$ ” and “A photo of a $\{class\}$ ” (step 1 of Algorithm 1). To generate a concrete prompt (step 7 of Algorithm 1), the placeholder $\{class\}$ is

replaced with an actual class name, *e.g.* *dog*, *car*, etc. According to step 6, about half of the images of each class are generated using the first prompt, while the other half using the second prompt. This

prompt variation is supposed to induce a higher variability in the generative process, thus obtaining more diverse images.

Active learning. An important factor for achieving a high accuracy while having a constraint on the number of API calls is the choice of samples to pass through the black-box model. Not only should the chosen images be representative, but they should also be diverse to improve the generalization capacity of the student. To this extent, we propose an active learning methodology to select representative and diverse samples to be inferred by the teacher.

At any iteration of the active learning procedure, we compute the latent vectors of all samples from the proxy subset X'' , given by the student (steps 14-15 of Algorithm 1). Based on the labels assigned by the student, we cluster the samples into classes and compute the centroid of each class (steps 16-17 of Algorithm 1). Next, we employ a sampling strategy that promotes the selection of examples closer to the centroids (steps 18-21 of Algorithm 1). The idea behind our strategy is to demote the selection of outliers, since these are more likely to be mislabeled by the teacher. We exploit the distance from each latent vector to its nearest centroid in a Radial Basis Function to compute the probability of sampling the corresponding image, as follows:

$$p_i = \exp\left(-\frac{\Delta(\bar{S}(x_i''), \mu_c)}{2 \cdot \sigma^2}\right), \quad (2)$$

where μ_c is the closest centroid to $\bar{S}(x_i'')$, and σ is a hyperparameter that controls the importance of the proximity to the nearest centroid. We select a uniformly distributed number of samples from each cluster to ensure the diversity of samples. The selected samples are given as input to the teacher, which returns a soft or hard label that is stored in Y' (steps 25-26 of Algorithm 1). We use the subset labeled by the teacher, denoted as X' , to train the student until convergence (steps 28-30 of Algorithm 1).

Self-paced learning. Since we assume that there is a limit imposed on the number of API calls, we can only retrieve class labels from the black-box model for only a fraction of our proxy data set. Hence, we have a large subset X'' of data samples that have not been labeled by the teacher. We propose a self-paced knowledge distillation procedure, in which we leverage the unlabeled data to improve the student. Our self-paced learning procedure gradually assigns labels to the remaining data using a nearest neighbor procedure applied in the latent embedding space learned by the student (steps 32-37 of Algorithm 1). More precisely, we operate in the latent space of the layer right before the flattening operation or the global average pooling layer, depending on the backbone architecture of our student. Let \bar{S} denote the latent space encoder. The first step of the proposed self-paced learning method is to pass each example from the annotated proxy subset X' through the student model and store the latent vectors and the labels assigned by the student. Then, for each unlabeled image $x_i'' \in X''$, we gather its corresponding latent representation and search for the closest k samples from the annotated training set (step 36 of Algorithm 1). To compute the distance in the latent space between two samples $x_i'' \in X''$ and $x_j' \in X'$, we consider two alternative metrics, namely the Euclidean

distance and the cosine distance. The latter is computed as follows:

$$\Delta_{\cos}(\bar{S}(x_i''), \bar{S}(x_j')) = 1 - \frac{\langle \bar{S}(x_i''), \bar{S}(x_j') \rangle}{\|\bar{S}(x_i'')\| \cdot \|\bar{S}(x_j')\|}, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product, and \bar{S} is the student encoder. Then, the label assigned to the sample x_i'' is inferred from the labels of its k nearest neighbors (step 37 of Algorithm 1). For the label assignment step, we suggest two schemes, depending on the output type received from the teacher. If the teacher provides soft labels, the resulting class distribution of image x_i'' is computed as a weighted average of the soft labels, where the weight of a sample x_j' is inversely proportional to the distance between x_i'' and x_j' . If the black-box model returns hard labels, we adopt a voting scheme based on plurality (majority) voting, where the distances between x_i'' and its neighbors are used to break ties. Finally, the student model is trained on the whole proxy data $X' \cup X''$ (steps 38-41 of Algorithm 1).

The active self-paced learning procedure is executed for a number of $r = n/s$ steps, until the API limit n is reached. We note that the latent space of the student changes during training. Hence, the latent vectors are computed at every step of the active learning procedure, to ensure that the sample selection procedure is on par with the current state of the student. During our experiments, we set $r = 3$, except for the one-shot and two-shot experiments, where r is constrained to $r = 1$ and $r = 2$, respectively.

4 EXPERIMENTS

4.1 Experimental Setup

Data sets. We conduct experiments on two image data sets, namely CIFAR-10 [23] and Food-101 [6]. CIFAR-10 is a data set of 50,000 training images and 10,000 test images, representing objects of 10 categories. Each image has a resolution of 32×32 pixels. Food-101 [6] is a data set containing images of 101 food categories. Each image has a resolution of 224×224 pixels. The original split contains 75,750 training images and 25,250 test images. These data set choices are aimed at testing the model stealing frameworks in distinct settings, comprising both low-resolution and high-resolution images, as well as a small and a large number of classes. The training sets are only used to train the black-box teachers. In contrast, the copy models are trained on generated proxy data.

Diffusion models. We generate two proxy data sets for CIFAR-10, one with Stable Diffusion v2 [40] and one with GLIDE [30]. The generated images are resized to match the input size of 32×32 pixels, as required by the black-box teacher. For each proxy data set, we generate 5,000 images per class. In Figure 2, we present one generated sample per class from each proxy data set.

For Food-101, we generate a proxy data set with 1000 images per class, using Stable Diffusion v2. We illustrate some randomly chosen synthetic images from this data set in Figure 3.

Although we can easily generate many more proxy images, we choose to limit the number of generated images in each proxy data set to the number of samples available in the original CIFAR-10 and Food-101 data sets. For each proxy data set, we keep 15% of the generated images for validation purposes.

Teacher and student models. In our experiments, we employ well-known model architectures that are fundamental to research,



Figure 2: Samples of generated images by GLIDE [30] (top row) and Stable Diffusion [40] (bottom row) for the CIFAR-10 classes.



Figure 3: Samples of generated images by Stable Diffusion [40] for some of the Food-101 classes.

as this facilitates comparison with other baselines. For the black-box models, we use two architectures: AlexNet [24] and ResNet-50 [16]. Following previous research on model stealing [1, 7], we consider lighter student architectures. For the AlexNet teacher, the student is Half-AlexNet, an architecture where the number of convolutional filters and the number of neurons in fully-connected layers are reduced by 50%. For the ResNet-50 teacher, the corresponding student is ResNet-18 [16].

Baselines. We compare our approach with two state-of-the-art model stealing methods [7, 35]. The first baseline is Black-Box Ripper [7], a framework that employs a generative model in order to create proxy data, but the framework is rather focused on achieving a high accuracy, irrespective of the number of API calls.

Knockoff Nets [35] represent our second baseline. Aside from their relevance in the model stealing research, Knockoff Nets have a similar focus to our own, namely to optimize the number of teacher

Table 1: Optimal hyperparameters of the student models for the CIFAR-10 data set.

Student	Diffusion model	Learning rate	Step size	γ
Half-AlexNet	GLIDE	$9 \cdot 10^{-4}$	20	0.95
ResNet-18	GLIDE	$9 \cdot 10^{-4}$	20	0.95
Half-AlexNet	Stable Diffusion	$6 \cdot 10^{-4}$	20	0.95
ResNet-18	Stable Diffusion	10^{-4}	30	0.9

Table 2: Optimal hyperparameters of the student models for the Food-101 data set.

Student	Diffusion model	Learning rate	Step size	γ
Half-AlexNet	Stable Diffusion	$7 \cdot 10^{-5}$	10	0.95
ResNet-18	Stable Diffusion	10^{-4}	30	0.95

(or victim) passes. Following Orekondy *et al.* [35], we use CIFAR-100 as proxy data for Knockoff Nets, when the evaluation is performed on CIFAR-10. Similarly, we use ImageNet-200 [41] (a subset of 200 classes from ImageNet) as proxy data for Food-101.

For a fair comparison, we impose the same limit on the number of API calls for all frameworks. Moreover, we use the same teacher and student architectures for all frameworks. Hence, the reported accuracy rates reflect the performance levels of the training frameworks, namely Black-Box Ripper [7], Knockoff Nets [35], and ASPKD (ours).

Hyperparameters. Throughout the experiments, we employ the Adam optimizer [22] with a decaying learning rate scheduler. The hyperparameters for the teachers are tuned independently of the students, thus preserving the black-box nature of the teachers. In the case of CIFAR-10, the teachers are trained for 100 epochs with early stopping and a learning rate of $5 \cdot 10^{-4}$ on mini-batches of 64 samples, while the scheduler has a step size of 5 with $\gamma = 0.95$. For the experiments on Food-101, the teachers are trained for 100 epochs with a learning rate of 10^{-3} and a mini-batch size of 64. For the learning rate scheduler, the step size is 20 with $\gamma = 0.95$.

The student models are fine-tuned on 15% of the proxy data. The students are trained for 100 epochs with early stopping on mini-batches of 64 samples. As far as the active learning strategy is concerned, we set the value of σ in Eq. (2) to 17. The nearest neighbors algorithm in the self-paced learning method uses $k = 5$ neighbors. The optimal values for the other hyperparameters of the students on CIFAR-10 are reported in Table 1, and those on

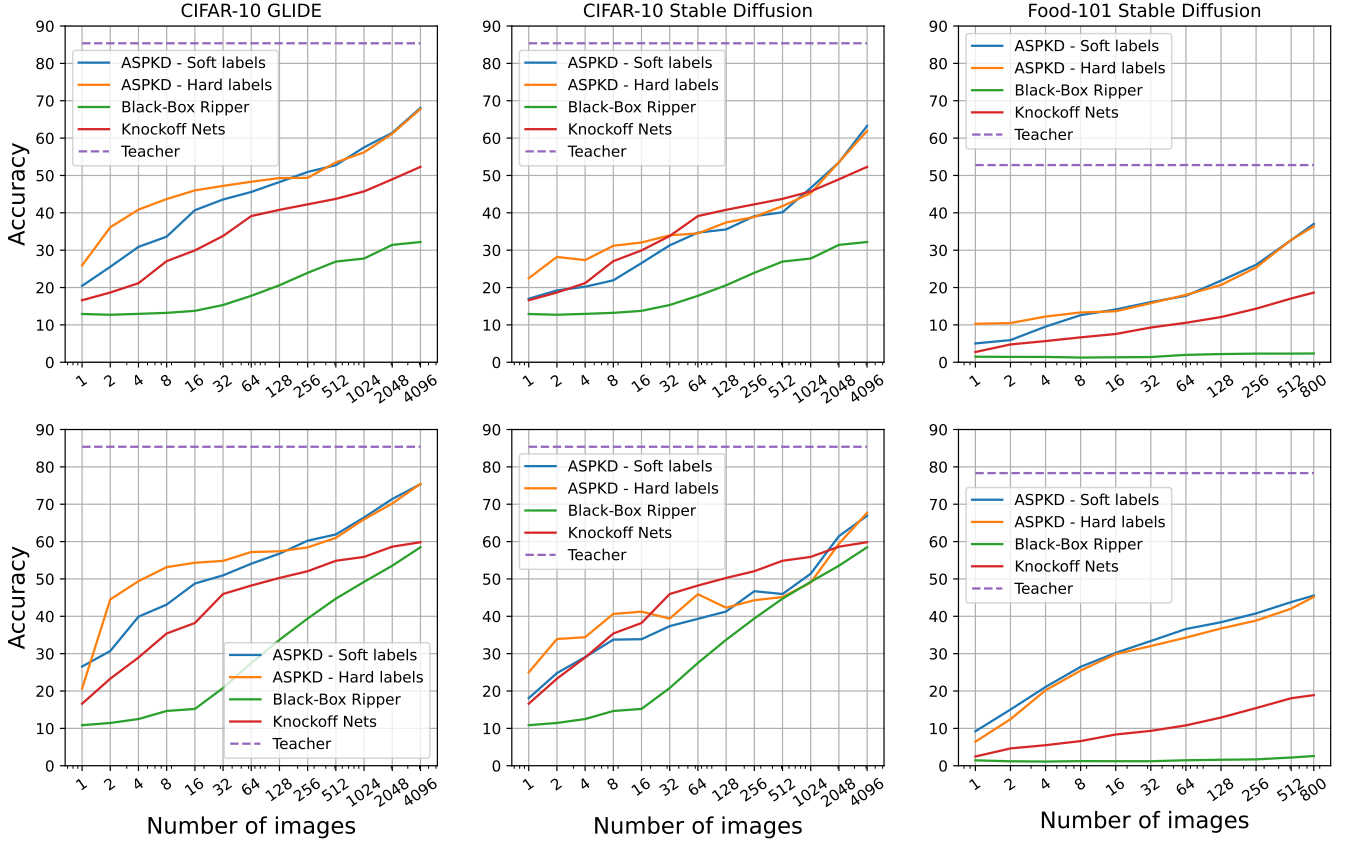


Figure 4: Empirical results for six experimental scenarios, where the maximum number of API calls per class takes values in the set $\{1, 2, 4, \dots, 4096\}$ for CIFAR-10, and the set $\{1, 2, 4, \dots, 512, 800\}$ for Food-101. The plots on the first two columns depict the results on CIFAR-10 [23], while the plots on the last column illustrate the results on Food-101 [6]. The proxy data for the plots on the first column is generated with GLIDE [30], while the proxy data for the other plots is generated with Stable Diffusion v2 [40]. For the plots on the top row, the student architecture is based on Half-AlexNet. For the plots on the bottom row, the student model is ResNet-18. We compare the results of ASPKD based on soft and hard labels with those of two state-of-the-art frameworks: Black-Box Ripper [7] and Knockoff Nets [35]. For reference, the accuracy rate of the corresponding teacher model is added to each plot. For each experiment, we report the average performance computed over 5 runs with each model. Best viewed in color.

Food-101 in Table 2. We present results with two versions for our framework: one that learns from hard teacher labels, and one that learns from soft teacher labels. For ASPKD based on hard labels, we use the Euclidean distance to find the nearest neighbors during self-paced learning. For ASPKD based on soft labels, we use the cosine distance defined in Eq. (3).

Evaluation. All models are evaluated on the official test sets of CIFAR-10 and Food-101. For the teacher models, we report the classification accuracy with respect to the ground-truth labels. Since the goal of the student models is to replicate the teachers, we evaluate each student in terms of the classification accuracy with respect to the labels predicted by its teacher. For each experiment, we report the average performance computed over 5 runs with each model.

4.2 Results

Main results. On CIFAR-10, we have four evaluation scenarios, since there are two teacher-student pairs and two diffusion models. On Food-101, we have two evaluation scenarios, as we use the

same teacher-student pairs, but only one diffusion model. In total, we compare our framework (ASPKD) with Black-Box Ripper [7] and Knockoff Nets [35] in six scenarios. The maximum number of API calls per class takes values in the set $\{1, 2, 4, \dots, 4096\}$ for the four CIFAR-10 scenarios, and the set $\{1, 2, 4, \dots, 512, 800\}$ for the two Food-101 scenarios. The corresponding results are presented in Figure 4. The synthetic training data is generated using GLIDE for the plots on the first column, and Stable Diffusion for the rest. For the plots on the first row, the teacher-student pair is represented by AlexNet→Half-AlexNet. For the second row, the teacher-student pair is ResNet-50→ResNet-18. In each plot, we present results with two versions for ASPKD, corresponding to the type of labels returned by the teacher, soft or hard.

When compared to Black-Box Ripper [7], our framework (ASPKD) obtains significantly better results in all six evaluation scenarios, regardless of the maximum number of API calls. In four evaluation scenarios (illustrated on the first and third columns in Figure 4),

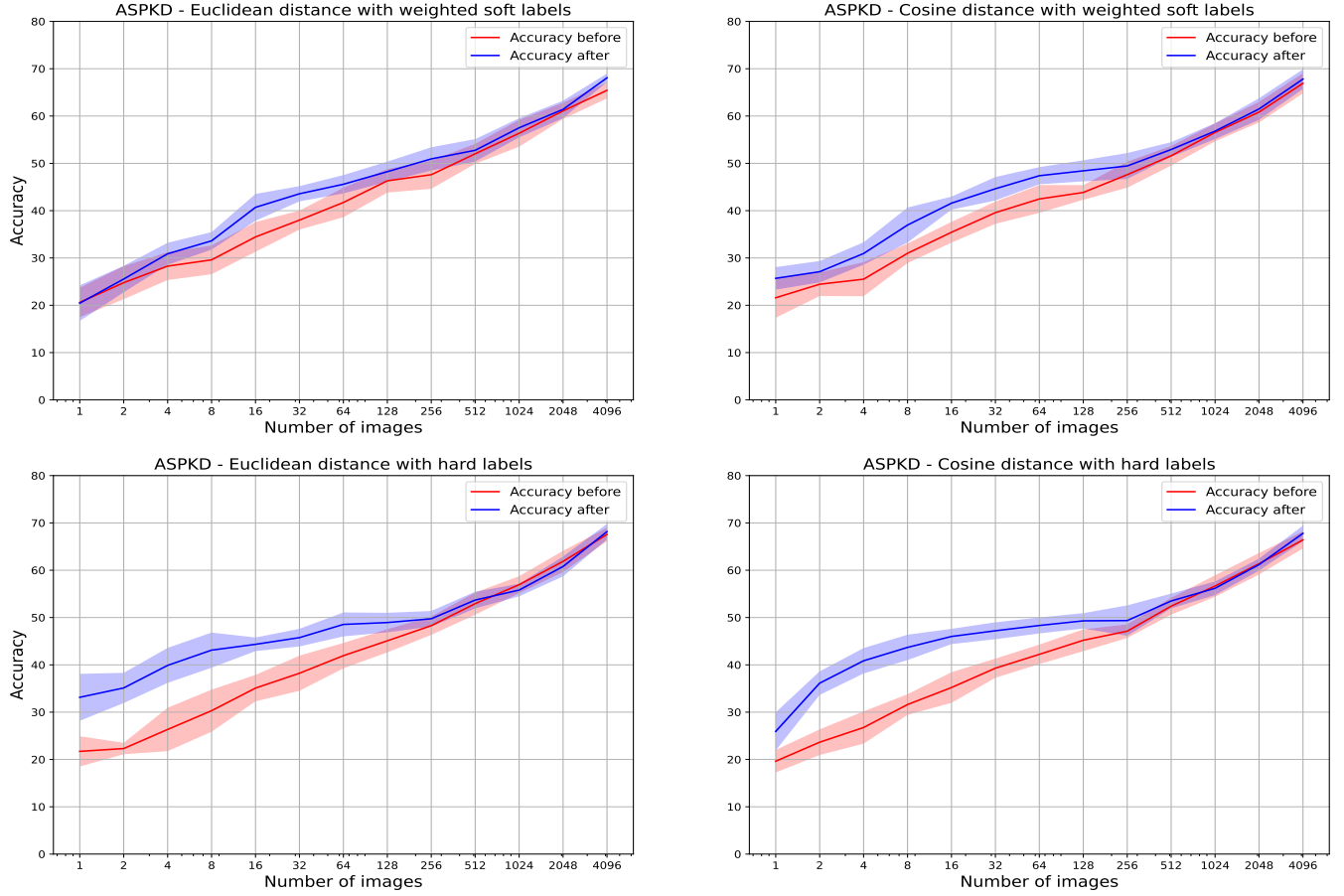


Figure 5: Accuracy rates before and after introducing our self-paced learning strategy. The comparison is carried out for four combinations of distance functions (Euclidean or cosine) and label types (hard or soft). For each experiment, we present the mean accuracy and the standard deviation over 5 runs. The test data is CIFAR-10, and the synthetic training images are generated by GLIDE [30]. The teacher is AlexNet, and the student is Half-AlexNet. Best viewed in color.

both ASPKD versions outperform Knockoff Nets [35] by considerable margins. For the other two scenarios, where the proxy data for CIFAR-10 is generated with Stable Diffusion, Knockoff Nets [35] temporarily surpass ASPKD, when the number of API calls per class ranges between 32 and 1024. ASPKD based on hard labels yields better results than Knockoff Nets in the more challenging few-call settings, namely when the maximum number of API calls per class is below 16. As the number of API calls per class increases, the two ASPKD versions register faster performance gains, recovering the temporary performance gap and even outperforming Knockoff Nets when the number of API calls per class is at least 2048. Considering the bigger picture, we conclude that ASPKD leads to generally better results, surpassing both Black-Box Ripper [7] and Knockoff Nets [35] in most evaluation scenarios.

Ablation studies. In order to demonstrate the capability of the individual stages of the proposed method, we carried out several ablation studies. To demonstrate the efficiency of our self-paced learning scheme, we conduct an analysis of the performance before and after introducing our self-paced strategy, considering the four possible combinations of distance functions (Euclidean or cosine)

and teacher labels (hard or soft). The corresponding results are illustrated in Figure 5. A clear pattern emerges when analyzing the four plots, specifically that self-paced learning brings considerable performance gains when the number of API calls per class is below 512. The improvements are usually higher for ASPKD based on hard labels. In conclusion, the empirical evidence indicates that our self-paced learning strategy plays a key role in the few-call model stealing scenarios.

In Table 3, we present the performance impact caused by alternatively and jointly introducing the active learning and the self-paced learning strategies, respectively. When we separately introduce the active learning and the self-paced learning strategies, we observe that each strategy brings significant gains in the majority of cases. Interestingly, when the number of API calls per class is below or equal to 128, we notice even higher gains when both strategies are jointly introduced. In summary, the results show the benefits of both strategies.

One of the dangers of self-paced learning is degrading the performance because of the amount of noise in the pseudo-labels. Given the uncertainty of the pseudo-labeling process, we next analyze the

Table 3: Accuracy rates of the Half-AlexNet student based on various training procedures. The vanilla procedure is based on training the student on proxy data with teacher labels in a conventional way. Next, we show the impact of separately and jointly introducing active learning and self-paced learning (based on cosine distance and soft labels), respectively. For each experiment, we present the mean accuracy and the standard deviation over 5 runs. The results are reported on CIFAR-10, while the proxy training data is generated by GLIDE [30].

#Samples per class	Vanilla	+ Active learning	+ Self-paced learning	+ Active & self-paced learning
1	19.4±2.1	20.7±3.6	25.0±2.7	25.9±4.0
2	25.9±3.7	26.5±2.0	25.6±1.1	36.1±2.5
4	27.7±2.6	29.9±1.2	28.5±0.5	40.9±2.7
8	29.4±2.0	33.0±2.0	35.3±1.7	43.7±2.7
16	36.3±3.5	41.8±3.0	40.8±1.1	46.0±1.6
32	39.2±2.1	43.9±1.3	46.4±1.7	47.2±1.8
64	43.6±1.4	46.0±2.7	45.8±1.2	48.3±1.7
128	46.6±1.8	48.8±1.5	48.0±1.6	49.3±1.6
256	47.7±3.2	50.3±1.9	50.0±1.5	49.4±3.2
512	52.5±2.1	53.6±1.9	53.6±0.8	53.5±1.6
1024	55.5±2.5	56.0±1.5	56.6±2.2	56.2±1.4
2048	60.3±1.7	61.5±2.1	63.1±0.9	61.2±1.3
4096	66.3±1.8	68.0±0.7	67.7±1.3	67.8±1.7

accuracy of the pseudo-labels with respect to the labels that would have been predicted by the teacher model. The corresponding results, which are shown in Figure 6, indicate that the quality of the pseudo-labels is consistently high, regardless of the number of API calls per class. This explains why our self-paced learning strategy works so well.

5 CONCLUSIONS

In this study, we explored the task of replicating the functionality of black-box machine learning models. We designed our method to be applicable in real-world scenarios, where there are several constraints, *i.e.* no access to the training set, no information about the architecture of the victim model or about its training process, as well as a cap on the number of permitted model calls. Our first contribution was to generate synthetic training data using a text-to-image diffusion model, allowing us to generate any class entity, while having a high diversity of images. Due to the limit imposed on the number of API calls, we introduced a self-paced learning method that assigns pseudo-labels for generated images that never get passed through the black-box teacher model. We also presented an active learning strategy that improves the process of selecting the proxy data to be labeled by the teacher. We carried out extensive experiments focusing on reducing the number of API calls, reporting results on various test cases based on multiple combinations of teacher-student architectures, distinct data sets, different diffusion models, and different output types given by the black-box model.

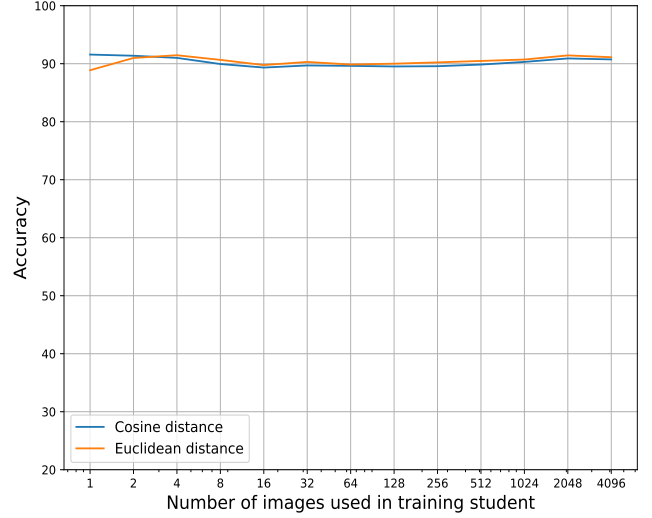


Figure 6: The accuracy of the pseudo-labels assigned during the self-paced learning process, with respect to the labels that would have been predicted by the teacher. The results are reported for the Half-AlexNet student on CIFAR-10, while the proxy training data is generated by GLIDE [30].

In the current surge of artificial intelligence solutions, our research aims to raise awareness of the exposure to model stealing attacks.

In future work, we aim to investigate methods to prevent few-call model stealing attacks.

6 ETHICAL CONSIDERATIONS

The possibility of launching model stealing attacks against machine learning models exposed via public APIs is a serious threat for companies releasing such models. Our results show that there is a high risk of stealing the intellectual property behind the released models, even when the access is restricted to just the output of the respective models. Indeed, we demonstrate how accessible it is to perform a model stealing attack, relying only on public information and accessible resources. We used two different open-source diffusion models to generate proxy data. Then, by querying the black-box model and obtaining its hard or soft labels for a limited number of images, our method can easily be applied to distill the knowledge of the black-box teacher into a copy model, which can later be used with no restrictions. This type of attack can be launched by any machine learning engineer, resulting in a potentially large number of intellectual property infringements. We consider that our work will inspire current researchers to continue on this track and work towards discovering methods of prevention against model stealing attacks. In this way, companies and individuals that publicly release models will benefit from enhanced security mechanisms.

REFERENCES

- [1] Sravanti Addepalli, Gaurav Kumar Nayak, Anirban Chakraborty, and Venkatesh Babu Radhakrishnan. 2020. DeGAN : Data-Enriching GAN for Retrieving Representative Samples from a Trained Classifier. In *Proceedings of AAAI*, Vol. 34. 3130–3137.
- [2] Tomer Amit, Eliya Nachmani, Tal Shaharabany, and Lior Wolf. 2021. SegDiff: Image Segmentation with Diffusion Probabilistic Models. *arXiv preprint arXiv:2112.00390* (2021).

- [3] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of CVPR*. 18208–18218.
- [4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrukov, and Artem Babenko. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *Proceedings of ICLR*.
- [5] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrđić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Proceedings of PKDD*. Springer, 387–402.
- [6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – Mining Discriminative Components with Random Forests. In *Proceedings of ECCV*. 446–461.
- [7] Antonio Bărbălu, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. 2020. Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. In *Proceedings of NeurIPS*, Vol. 33. 20120–20129.
- [8] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. 2020. Exploring connections between active learning and model extraction. In *Proceedings of USENIX*. 1309–1326.
- [9] Jacson Rodrigues Correia-Silva, Rodrigo F. Berriel, Claudine Badue, Alberto F. de Souza, and Thiago Oliveira-Santos. 2018. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data. In *Proceedings of IJCNN*. 1–8.
- [10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 9 (2023), 10850–10869.
- [11] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Proceedings of NeurIPS*, Vol. 34. 8780–8794.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of ICLR*.
- [13] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of AISTATS*. 249–256.
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NeurIPS*, Vol. 27. 2672–2680.
- [15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*. 770–778.
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of NeurIPS*, Vol. 33. 6840–6851.
- [18] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. 2020. High accuracy and high fidelity extraction of neural networks. In *Proceedings of USENIX*. 1345–1362.
- [19] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. 2019. PRADA: protecting against DNN model stealing attacks. In *Proceedings of EuroS&P*. 512–527.
- [20] Sanjay Kariyappa, Atul Prakash, and Moinuddin K. Qureshi. 2021. MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation. In *Proceedings of CVPR*. 13814–13823.
- [21] Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. 2018. Model Extraction Warning in MLaaS Paradigm. In *Proceedings of ACSAC*. 371–380.
- [22] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of ICLR*.
- [23] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images*. Technical Report. University of Toronto.
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of NeurIPS*, Vol. 25.
- [25] Xinjing Liu, Zhuo Ma, Yang Liu, Zhan Qin, Junwei Zhang, and Zhuzhu Wang. 2022. SeInspect: Defending Model Stealing via Heterogeneous Semantic Inspection. In *Proceedings of ESORICS*. 610–630.
- [26] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *Proceedings of CVPR*. 11461–11471.
- [27] Paul Micaelli and Amos J. Storkey. 2019. Zero-shot knowledge transfer via adversarial belief matching. In *Proceedings of NeurIPS*, Vol. 32. 9551–9561.
- [28] Itay Mosafi, Eli Omid David, and Nathan S. Netanyahu. 2019. Stealing knowledge from protected deep neural networks using composite unlabeled data. In *Proceedings of IJCNN*. 1–8.
- [29] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. 2019. Zero-shot knowledge distillation in deep networks. In *Proceedings of ICML*. 4743–4751.
- [30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photo-realistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of ICML*. 16784–16804.
- [31] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of ICML*. 8162–8171.
- [32] Seong Joon Oh, Bernt Schiele, and Mario Fritz. 2019. Towards reverse-engineering black-box neural networks. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (2019), 121–144.
- [33] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. 2023. I know what you trained last summer: A survey on stealing machine learning models and defences. *Comput. Surveys* 55, 14s, Article 324 (2023).
- [34] OpenAI. 2022. ChatGPT: A Conversational Language Model. <https://openai.com/research/chatgpt>.
- [35] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff Nets: Stealing functionality of black-box models. In *Proceedings of CVPR*. 4954–4963.
- [36] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. 2019. A framework for the extraction of deep neural networks by leveraging public data. *arXiv preprint arXiv:1905.09165* (2019).
- [37] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of ASIACCS*. 506–519.
- [38] Li Pengcheng, Jinfeng Yi, and Lijun Zhang. 2018. Query-Efficient Black-Box Attack by Active Learning. In *Proceedings of ICDM*. 1200–1205.
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*. 8748–8763.
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of CVPR*. 10684–10695.
- [41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proceedings of NeurIPS*, Vol. 35. 36479–36494.
- [43] Sunandini Sanyal, Sravanti Addepalli, and R. Venkatesh Babu. 2022. Towards data-free model stealing in a hard label setting. In *Proceedings of CVPR*. 15284–15293.
- [44] Yi Shi, Yalin Sagduyu, and Alexander Grushin. 2017. How to steal a machine learning classifier with deep learning. In *Proceedings of HST*. 1–5.
- [45] Yi Shi, Yalin E. Sagduyu, Kemal Davaslioglu, and Jason H. Li. 2018. Active deep learning attacks under strict rate limitations for online API calls. In *Proceedings of HST*. 1–6.
- [46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of ICML*. 2256–2265.
- [47] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proceedings of ICLR*.
- [48] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *Proceedings of USENIX*. 601–618.
- [49] Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing hyperparameters in machine learning. In *Proceedings of SP*. IEEE, 36–52.
- [50] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. 2022. Diffusion Models for Medical Anomaly Detection. In *Proceedings of MICCAI*. 35–45.
- [51] Yi Xie, Mengdie Huang, Xiaoyu Zhang, Changyu Dong, Willy Susilo, and Xiaofeng Chen. 2022. GAME: Generative-Based Adaptive Model Extraction Attack. In *Proceedings of ESORICS*. 570–588.
- [52] Haonan Yan, Xiaoguang Li, Hui Li, Jiamin Li, Wenhui Sun, and Fenghua Li. 2022. Monitoring-Based Differential Privacy Mechanism Against Query Flooding-Based Model Extraction Attack. *IEEE Transactions on Dependable and Secure Computing* 19, 4 (2022), 2680–2694.
- [53] Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. 2020. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion. In *Proceedings of CVPR*. 8715–8724.
- [54] Zhanyuan Zhang, Yizheng Chen, and David Wagner. 2021. SEAT: Similarity Encoder by Adversarial Training for Detecting Model Extraction Attack Queries. In *Proceedings of AISec*. 37–48.