KOEnsAttack: Towards Efficient Data-Free Black-Box Adversarial Attacks via Knowledge-Orthogonalized Substitute Ensembles

Chaoyong Yang[1,2] Jia-Li Yin[1,2*] Bin Chen[1,2] Zhaozhe Hu[1,2] Xiaolei Liu[3] Wei Lin[4*]
[1]Fujian Province Key Laboratory of Information Security and Network Systems, Fuzhou, China
[2]Fuzhou University [3]China Academy of Engineering Physics [4]Fujian University of Technology
{231020055, jlyin, 231010004, 241010027}@fzu.edu.cn, luxaole@gmail.com, wlin@fjut.edu.cn

Abstract

Data-free black-box attacks aim to attack a model without access to either the model parameters or training data. Existing methods use a generator to synthesize training samples and then train a substitute model to imitate the victim model. The adversarial examples (AEs) are finally generated using the substitute model to transfer to the victim model. To this end, how to generate diverse training samples for substitute model training and improve the transferability of AEs from the substitute model to victim model become the core challenges. In this paper, we propose a Knowledge-Orthogonalized Ensemble Attack, dubbed KOEnsAttack, to accomplish these two goals. We first use dual networks as the ensemble substitute model, and then propose a sample hardness enhancement to transform the samples from the generator into hard samples that exist in the controversial regions of the dual models for promoting the sample diversity. Next, during the substitute model training, we design a knowledge orthogonalization module to guide the dual networks in learning complementary and useful information from the victim model, thereby enhancing the transferability of adversarial samples generated on the final ensemble model. Extensive experiments on several datasets are conducted to evaluate the effectiveness of our method. The results show that the proposed method can achieve superior performance compared with the state-of-the-art competitors.

1. Introduction

Deep Neural Networks (DNNs) have emerged as the dominant framework across various applications [4, 41] owing to their demonstrated capacity for learning hierarchical representations from complex data structures. However, recent studies [16, 28, 33–35] have exposed an inherent susceptibility of DNNs to adversarial examples (AEs), wherein strategically crafted perturbations—often imperceptible to human observers—can systematically deceive model predictions. This fundamental vulnerability raises critical concerns regarding the operational security of DNN-based systems in real-world deployments. Consequently, the machine learning community has intensified efforts to address this dichotomy through two complementary research thrusts, namely adversarial attack and defense, for better improving the robustness of networks.

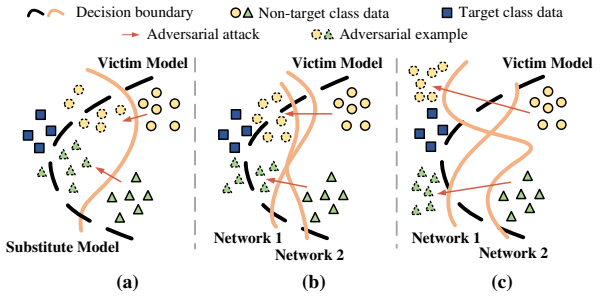

Figure 1. Comparison of different generator-based data-free black-box attacks. (a) A substitute model is trained to imitate the victim model, however, the reasoning gap between the substitute model and the victim model can be large due to the distribution shift in the synthesized samples; (b) Recent methods extend dual networks as the substitute model to minimize the gap, but the sample space are not fully explored which leads to under-optimal results; (c) We propose the knowledge-orthogonalized ensembles to reinforce the generalization of the substitute model which can effectively improve the attack performance.

Typically, adversarial attacks can be categorized as white-box attacks [1, 18, 19, 28] and black-box attacks [2, 15, 32] according to the level of access to the victim model. Existing attacks can achieve near-optimal attack performance under the white-box settings via reversing the gradients as the perturbations; however, they would lead to trivial solutions under the black-box settings where only the input-output feedback of the victim model can be accessed. Modern black-box attacks typically leverage the transferability of adversarial perturbations, crafting AEs on

*Corresponding authors.

a pre-trained substitute model through specific operations to threaten the target model. However, in more realistic settings, the training data can also be unavailable, which limits the training of substitute models, making the adversarial attack still a struggle.

To address the "data-free" issue, a number of works [23, 25, 27, 30, 40] additionally utilize a well-designed generator to synthesize training samples for training the substitute model. However, due to the distribution shift in the synthesized training samples, the reasoning gap between the substitute model and target model can be large and further hinders the transferability of AEs, as shown in Figure 1 (a). To minimize the gap, recent studies [23] propose the ensemble substitute model that is composed of dual networks, and then a disagreement loss is employed on the generator to produce hard samples that can maximize the disagreement between these two networks. These hard samples provide the driving force to make the decision boundary of at least one of the dual networks to align with the victim model in each training epoch since they are misclassified by at least one network in the ensemble, as shown in Figure 1 (b). While remarkable progress can be achieved, we notice that there still exist several limitations: First, the hard samples in the controversial regions of the input space can play a critical role in substitute model training because they can not only help minimize the reasoning gap but also reduce the query times since they are more effective than the easy data. However, the loss optimization in the existing methods cannot fully explore the input space, especially when combined with other objectives. Second, for the substitute model training, existing methods only use the discrimination loss to imitate the victim model, neglecting the generalization of the substitute model which potentially impedes the transferability of the generated AEs.

In this paper, we design the Knowledge-Orthogonalized Ensemble Attack, dubbed KOEnsAttack, for efficient data-free black-box adversarial attacks. Specifically, we follow the previous data-free adversarial attack pipeline where a generator is employed to generate training samples and a substitute model is trained on these samples with the query feedback of the victim model. The AEs are finally generated by white-box attacks on the trained substitute model. Instead of employing disagreement loss on the generator, we propose to obtain the hard samples in a more direct and effective way: iteratively adding the reversed gradients of the discrepancy loss between two sub-networks to the original samples, so that the generated samples can be differently classified in the dual networks. By such transformation, the samples are all hard samples for substitute model training, which can effectively push the similarity to the victim model and reduce the query times during training. Meanwhile, we further design a knowledge orthogonalization module to force the dual networks to learn complementary knowledge from the same black-box model, reinforcing the generalization of the substitute model and finally improving the transferability of the generated AEs, as shown in Figure 1 (c). Our proposed method demonstrates 98.12% untargeted attack success rate on CIFAR-100 dataset within only 4M query times, surpassing the SOTA method DisGUIDE [23] by 1.02% and a 6M query reduction.

The main contributions of our work can be summarized as follows:

- We propose KOEnsAttack, a query-efficient black-box attack method that employs an ensemble substitute model trained on hard samples that exist in the controversial regions of the sub-networks in the ensemble.
- We introduce the sample hardness enhancement (SHE) for transforming the generated samples into hard samples that are beneficial for substitute model training and effectively reduce the black-box query budget.
- We further design the knowledge orthogonalization module (KOM) to improve the generalization of the ensemble substitute model, which consequently promotes the transferability of the generated AEs.
- Our empirical evaluations on various datasets under both untargeted and targeted attacks demonstrate that the proposed KOEnsAttack can achieve SOTA attack performance with a large margin and significantly reduce the query budgets during the substitute training process.

2. Related Work

Adversarial Attacks. Since Szegedy *et al.* [28] first illustrated the vulnerability of DNNs to adversarial examples, extensive studies [6, 9, 17, 18, 22, 36, 39] have been conducted to focus on adversarial attacks for misleading the well-trained DNN models. In general, these methods can be divided into white-box and black-box attacks according to whether the attackers have access to the parameters or architectures of the target model. By leveraging the reversed gradients, white-box attacks can achieve high attack success rates; however, the attack performance would degrade severely under the black-box settings since attackers only access the simple output of the victim model. A line of methods [3, 8, 31] utilizes inputs query feedback to guide the generation of adversarial perturbations, while another line [2, 5, 11, 20] aims to improve the transferability of adversarial examples crafted from a substitute model to unknown models. Despite the achievements, in the more realistic scenarios, attackers can access neither the training data nor the target model, *i.e.*, data-free black-box setting, bringing extra challenges in substitute model training.

Data-Free Black-Box Attacks. Exposed to the challenge of data-free and black-box settings, recent works [23, 27, 30] proposed to employ a generator to generate training samples and then use these samples to train a substi-

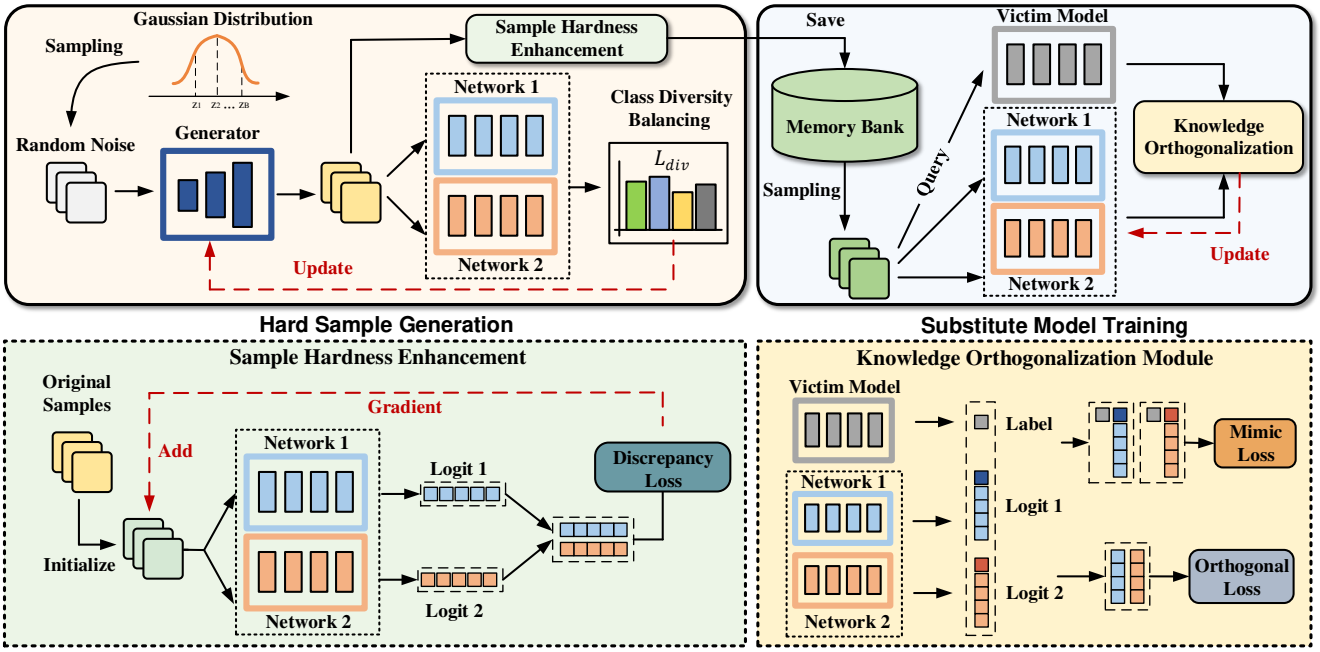

Figure 2. Illustration of our KOEnsAttack pipeline. A generator fed with random noise is first trained to produce synthetic data, followed by the sample hardness enhancement (SHE) strategy to transform the samples into hard samples. Next, the ensemble substitute model is trained via a knowledge orthogonalization module (KOM) to encourage the learning of valuable and complementary information from the victim model.

tute model for generating transferable adversarial examples. DaST [40] is the first work to utilize a generator with noise as input to synthesize data for querying the target model and training the substitute model. DFME [30] and MAZE [12] further estimated the gradients of the victim model using black-box gradient estimators to force the similarity between the substitute model and target model. To improve the synthesized data quality, Sanyal *et al.* [25] employed a generative adversarial network (GAN) framework by introducing additional datasets as the proxy data to improve the data quality. Rethinking the convergence failure and model collapse of previous methods between generator and substitute model, IDEAL [38] designed a powerful black-box attack framework that two players are no longer forced to directly compete in min-max game. Besides, DisGUIDE [23] proposed to maximize the disagreement loss between two surrogate models to force the generator to produce more query-valuable synthetic samples. More recently, STDatav2 [27] introduces the joint-data optimization that leverages both synthesized and proxy data and develops a self-conditional data synthesis framework for improving data diversity. While remarkable progress has been achieved, the loss-based optimization in sample synthesis of these existing methods cannot fully explore the sample space for substitute model training and the generalization ability of the substitute model should also be considered for transferability improvement.

3. Method

3.1. Overview

Our goal is to attack a model in a data-free and black-box scenario. Given a black-box victim model $\mathcal{V}$ without any knowledge (*i.e.*, model structure, parameters, and training data), we first introduce a generator $\mathcal{G}$ to synthesize training samples X. And then a transparent white-box substitute model $\mathcal{S}$ is trained on X to imitate $\mathcal{V}$ for generating AEs that can be transferred to attack $\mathcal{V}$. Note that we follow the most realistic scenario that only the label output of $\mathcal{V}$ is accessible, *i.e.*, hard label setting.

We present KOEnsAttack, a generator-based framework with an ensemble substitute model, to generate transferable AEs for the victim model. Our KOEnsAttack boosts the effective and efficient training of the substitute model by transforming the generated samples into hard samples as well as introducing knowledge orthogonalization into the ensemble for improving the generalization. An overview of our approach is shown in Fig. 2. It mainly consists of two phases: (1) Hard Sample Generation and (2) Substitute Model Training. The generator $\mathcal{G}$ is first trained to synthesize the training samples that are used for the substitute model training. We utilize the sample hardness en-

hancement to transform the output samples into hard samples. Then in phase 2, the substitute model $\mathcal{S}$ is trained on the hard samples with the query feedback from the victim model. Note that $\mathcal{S}$ is an ensemble of dual networks, denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$, where the final ensemble output is an average soft-vote of the dual networks.

3.2. Hard Sample Generation

In the first phase, the generator $\mathcal{G}$ takes noise Z as input and output samples X. To be formal, given the noise Z from Gaussian prior, $Z \sim \mathcal{N}(0,1)$, the generator $\mathcal{G}$ is utilized to map Z to the synthetic samples X for querying victim model $\mathcal{V}$. The synthetic samples X are expected to be evenly distributed in the sample space and exist in the controversial regions of the dual networks. Thus we first employ a class diversity loss to train the generator $\mathcal{G}$, instead of combining with a disagreement loss, we propose to directly transform the generated samples into hard samples using the sample hardness enhancement module.

Class Diversity Loss. We employ the class diversity loss to balance the generated data distribution and promote the data diversity. Specifically, we use information entropy to measure the confusion degree of the synthetic data, ensuring the diversity of the generated sample categories. Given a batch with batch size B of synthetic samples $X = \{x_i\}_{i=1}^{B} = \mathcal{G}(Z)$, we first compute the corresponding prediction of the substitute model as:

$$P_{ens}(x_i) = \frac{1}{2}(softmax(\mathcal{S}_1(x_i)) + softmax(\mathcal{S}_2(x_i))), \tag{1}$$

and then the generator $\mathcal{G}$ is optimized as the following loss:

$$\mathcal{L}_{\mathcal{G}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} P_{ens}^{k}(x_i)\log(P_{ens}^{k}(x_i)), \tag{2}$$

where $P_{ens}^{k}(x_i)$ is the k-th element of $P_{ens}(x_i)$, *i.e.*, the ensemble prediction score of the k-th class.

Sample Hardness Enhancement. Next, after updating $\mathcal{G}$ using $\mathcal{L}_{\mathcal{G}}$, we can obtain the samples $X = \mathcal{G}(Z)$. We propose to further move these samples to the controversial regions of the dual networks in the substitute model. Inspired by the gradient-based attack which has high success in making samples cross the decision boundary, we propose to iteratively reverse the gradient of the discrepancy loss between the dual networks on the samples. Specifically, the discrepancy loss between the dual networks can be first computed using the cosine similarity as:

$$\mathcal{L}_{dis} = \frac{1}{B}\sum_{i=1}^{B}\cos(\mathcal{S}_1(x_i), \mathcal{S}_2(x_i)), \tag{3}$$

where $\cos(\cdot)$ denotes the cosine similarity between two prediction outputs. This constraint aims to explore the sample space in search of synthetic samples that maximize the predictive discrepancy between the dual networks. Then, we optimize the original synthetic samples by directly maximizing the $\mathcal{L}_{dis}$ as follows:

$$\overline{x}_i^{t+1} = \overline{x}_i^{t} + \alpha \cdot \text{sign}(\nabla_{\overline{x}_i^t}\mathcal{L}_{dis}), \tag{4}$$

where $\nabla_{\overline{x}_i^t}$ denotes the gradient of the loss function $\mathcal{L}_{dis}$ w.r.t. the hard sample $\overline{x}_i^t$ in the dual networks $\mathcal{S}_1$ and $\mathcal{S}_2$, and α is the step size. The final hard sample $\overline{x}_i$ is obtained by $\overline{x}_i^T$, where T is the number of iterative steps.

Algorithm 1 Proposed KOEnsAttack Method.

Input: Random noise $Z \sim \mathcal{N}(0,1)$, victim model $\mathcal{V}$, substitute model $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2)$, generator $\mathcal{G}$, epochs E, total iters N, replay iters R, SHE iterative steps T, and memory bank $\mathcal{D}$.
Output: $(\mathcal{S}_1, \mathcal{S}_2)$

for $i = 1$ to E **do**
  for $n = 1$ to N **do**
    // Hard Sample Generation
    Generate a batch of data $X \leftarrow \mathcal{G}(Z)$
    Compute class diversity loss $\mathcal{L}_{\mathcal{G}}$
    Update generator $\mathcal{G}$ using $\mathcal{L}_{\mathcal{G}}$
    Let $\overline{X}^0 = X$
    for $t = 1$ to T **do**
      $\overline{X}^{t+1} \leftarrow \overline{X}^{t} + \alpha \cdot \text{sign}(\nabla_{\overline{X}^t}\mathcal{L}_{dis})$.
    end for
    Save hard samples $\overline{X}$ to $\mathcal{D}$
  end for
  // Substitute Model Training
  for $r = 1$ to R **do**
    Sampling data $\overline{X}$ from $\mathcal{D}$
    Compute $\mathcal{L}_{ens}$ for $\mathcal{S}_1$ and $\mathcal{S}_2$
    Update $\mathcal{S}_1$ and $\mathcal{S}_2$
  end for
end for
return $(\mathcal{S}_1, \mathcal{S}_2)$

3.3. Substitute Model Training

After obtaining the desired hard synthetic data $\overline{X}$ in the phase 1, we can train the substitute model $\mathcal{S}$ by querying the victim model $\mathcal{V}$ in phase 2. Most of the previous works have only focused on how to train the substitute models to better mimic the behavior of victim models. This objective setting is actually sub-optimal. We believe that a new model optimization objective should be reset to directly focus on how to obtain a suitable substitute model on which adversarial samples with high transferability against the specific victim model can be easily generated.

Thus, for a particular victim model, how can we find the most suitable substitute model for generating black-box adversarial perturbations? Let us define $\mathcal{H}$ as a hypothesis set

that contains all classifiers that perform well on the specified classification task. We also reasonably assume that the black-box model h_b and two arbitrary substitute models h, h' satisfy $h_b, h, h' \in \mathcal{H}$. When h and h' have decision boundaries similar to h_b, and their decision boundaries are as orthogonal to each other as possible, any adversarial sample that successfully misleads h and h' is higly likely to successfully mislead h_b. In this way, the model obtained by integrating h and h' can generate AEs that are highly likely to transfer to the black-box h_b.

As analyzed above, we propose the knowledge orthogonalization module (KOM) to constrain the substitute training, encouraging the dual networks to learn complementary and useful knowledge from the query feedback of the victim model. The goal of KOM is to enhance the generalization and transferability of adversarial samples generated on the final surrogate ensemble model. Specially, the essence of knowledge orthogonalization lies in linking certain intuitive representations of the model to the abstract knowledge it embodies, and constraining the knowledge learned by multiple students through specific and effective loss terms. Inspired by related works [5, 23], we intend to perform the knowledge orthogonalization at the logit level of the surrogate model's output during the model training phase.

Knowledge Orthogonalization Module. Firstly, the prediction behavior should be similar to that of the victim model. This can be achieved using the loss as follows:

$$\mathcal{L}_{ce} = \sum_{i=1}^{N} \ell_{ce}(\mathcal{S}_1(\overline{x}_i), \mathcal{V}(\overline{x}_i)) + \ell_{ce}(\mathcal{S}_2(\overline{x}_i), \mathcal{V}(\overline{x}_i)), \quad (5)$$

where $\ell_{ce}(\cdot)$ is the cross-entropy loss. Please note that $\mathcal{V}(\cdot)$ denotes one-hot outputs of the victim model since only label output can be accessible. To encourage each network to learn useful information about different aspects of the black-box model, we further propose the orthogonal loss that measures the discrepancy between the non-target predicted logits of the dual networks:

$$\mathcal{L}_{ol} = \sum_{i=1}^{N} |\cos(M(\mathcal{S}_1(\overline{x}_i)), M(\mathcal{S}_2(\overline{x}_i)))| \quad (6)$$

where $M(\cdot)$ is the logits of non-target classes. The goal of $\mathcal{L}_{ol}$ is to enforce orthogonality among the non-target class logit vectors of multiple students, enabling them to learn richer and more complementary knowledge.

In summary, the total loss function in the KOM for the substitute ensemble model as follows:

$$\mathcal{L}_{ens} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{ol}, \quad (7)$$

where λ controls the weight value of $\mathcal{L}_{ol}$. Please note that during the substitute model training, the dual networks are updated in parallel. Finally, the adversarial examples are generated by applying classic gradient-based white-box attacks on the substitute model, where the aggregated gradients are used from the dual networks.

Furthermore, since querying new instances is expensive (with pay-per-query systems), it implies that we need to fully exploit the valuable information embedded in synthetic samples instead of only using the online-generated samples. At the same time, previous data-free works often fail to account for the distribution shift that occurs in synthetic data during the substitute model training. This oversight can lead to catastrophic forgetting, resulting in unsatisfactory performance of the final model. Therefore, we adopt the experience replay from existing work [12], aiming to better leverage available synthetic data to improve the training efficiency of the surrogate model. In particular, we use a memory bank to store all previously synthesized data optimized by the sample hardness enhancement. Then, after each substitute training step, we randomly select several batches of synthetic samples from the memory bank to consolidate and reinforce the previously learned dark knowledge. The algorithm is summarized in Algorithm 1.

4. Experiments

4.1. Experiment Setup

Datasets and Model Architectures. We consider four datasets (SVHN [21], CIFAR-10 [13], CIFAR-100 [13] and Tiny ImageNet [24]) which are commonly used in data-free black-box attack research to verify the effectiveness of our method. For the model architecture, we utilize the pre-trained ResNet-18 [10], VGG-19 [26], ResNet-34 [10], ResNet-50 [10], and ViT [7] as the victim models. The ensemble of VGG-13 [26] and Inception-v3 [29] is adopted as the default substitute model.

Compared Methods. We compare our method with various baselines: (1) Data-driven methods that require real data as the proxy, *i.e.*, DFMS [25], STDatav2 [27]; (2) Data-free methods that truly do not rely on natural data, *i.e.*, DFME [30], DFTA [37], IDEAL [38] and DisGUIDE [23].

Attack Methods and Evaluation Metrics. We use several classic white-box attack methods to generate AEs over the well-trained substitute model, including FGSM [9], PGD [18] and BIM [14]. The attack success rate (ASR) is calculated by N/N_{total} as the evaluation metric, where N and N_{total} are the number of AEs that can fool the attacked model and the total number of AEs, respectively.

Implementation Details. During training, we utilize Adam and SGD to train our generator and substitute models from Pytorch scratch, respectively. Following the hyperparameter settings used in DFME [30], the Adam optimizer is configured with an initial learning rate of 1×10^{-4} and a weight decay rate of 5×10^{-4}. Similarly, the initial learn-

Type	Settings		Victim Dataset	SVHN	CIFAR-10			CIFAR-100			Tiny ImageNet
	Hard Label	Proxy Data	Target Model	ResNet-18	VGG-19	ResNet-34	ViT	VGG-19	ResNet-34	ViT	ResNet-50
Targeted	✗	✗	DFME [30]	88.62	57.20	85.24	34.70	3.469	42.12	14.53	2.601
	✓	✓	DFMS [25]	89.72	87.67	43.21	12.40	2.819	8.413	4.947	0.292
	✓	✗	DFTA [37]	17.26	27.38	27.22	12.05	4.997	11.81	7.284	0.020
	✓	✗	IDEAL [38]	14.40	36.81	31.81	12.78	5.048	9.264	7.092	0.151
	✓	✗	DisGUIDE [23]	75.15	69.46	95.32	31.56	29.02	57.20	18.34	1.025
	✓	✓	STDatav2 [27]	86.10	95.13	91.48	44.23	18.58	43.37	14.18	0.121
	✓	✗	KOEnsAttack (Ours)	**92.07**	**95.44**	**97.82**	**51.79**	**42.87**	**60.48**	**32.35**	**15.22**
Untargeted	✗	✗	DFME [30]	95.86	95.20	97.14	64.17	81.88	95.21	71.74	57.58
	✓	✓	DFMS [25]	96.37	99.38	78.12	33.95	74.72	72.58	50.76	34.27
	✓	✗	DFTA [37]	41.60	57.17	52.90	33.70	68.68	73.34	55.86	8.479
	✓	✗	IDEAL [38]	33.83	71.55	64.12	33.16	70.27	71.63	55.60	24.67
	✓	✗	DisGUIDE [23]	89.46	95.50	99.60	61.23	92.35	95.98	74.55	39.35
	✓	✓	STDatav2 [27]	94.66	99.66	99.14	79.60	87.39	93.42	70.90	17.64
	✓	✗	KOEnsAttack (Ours)	**97.41**	**99.89**	**99.78**	**82.60**	**96.73**	**97.85**	**85.43**	**78.23**

Table 1. Comparison of our proposed KOEnsAttack with other competitors on four datasets. For fair comparison, we use PGD as the attack method. Bolded texts indicate the best results, while underlines represent the second-best results.

ing rate of the SGD optimizer with the same weight decay as above is set as 0.1. Meanwhile, we utilize learning rate scheduler with a multiplying factor of 0.3 at intervals specified by fractions $[40\%, 80\%]$ of the query budget as DisGUIDE [23]. The model is trained on NVIDIA GeForce GTX 3090 GPU. Additionally, we follow [25] to transform a fraction of the generated samples to grayscale for enhancing the class diversity of the synthetic dataset.

4.2. Main Results

On Attacking Different Victim Models. As shown in Table 1, we conduct extensive comparisons with competitors from several aspects, *i.e.*, the diverse datasets, various target models, and different attack scenarios. We can see that our method beats all competitors with a large margin in terms of both targeted and non-targeted attack scenarios, especially in the targeted attack scenario. Meanwhile, the comparisons between our method and other baselines, DFME [30], DFMS [25] and STDatav2 [27], are particularly more noteworthy. First, unlike other baseline methods, DFME uses the forward estimation approach to approximate the backpropagation gradients of the black-box model, which requires access to the probability outputs from the target model. Compared to the hard labels, soft labels provide richer class information, which often results in better black-box attack performance. As shown in Table 1, among all baseline methods, DFME achieves second-best performance on the SVHN dataset and best performance on the Tiny ImageNet dataset, respectively. Remarkably, even in the scenario where only hard-label predictions are accessible, our method still outperforms DFME. Next, DFMS and STDatav2 are the only two baseline methods that utilize external real data. Generally, the training efficiency of substitute models can be greatly improved with additional prior knowledge about the image distribution. As demonstrated in Table 1, DFMS and STDatav2 demonstrate strong performance on small-scale datasets, respectively. Excitingly, our KOEnsAttack, without relying on any real data prior, still achieves superior black-box attack performance compared to these two data-driven approaches.

Moreover, we show the curves about the attack success rate of targeted PGD attacks for all methods under different black-box query budgets across various datasets in Figure 3 (a)-(b). The results demonstrate that our method is significantly more efficient compared to other baseline approaches, substantially reducing the required black-box query budget while maintaining the high transferability of adversarial samples generated on the substitute model.

On the Combination with Different White-Box Attacks. We further compare our methods with competitors when combined with different white-box attacks on well-trained substitute models. As illustrated in Table 2, we apply three classic attacks to generate AEs over substitute models for attacking the black-box victim models on four datasets. Compared to other data-free black-box adversarial attacks, our KOEnsAttack method can efficiently acquire highly query-valuable synthetic samples through a few iterations of enhancement and introduces a novel optimization objective during the model training phase to effectively train the ensemble surrogate model. As a result, it achieves state-of-the-art black-box attack performance even under limited query budgets. Notably, it is well-known that FGSM, as the most basic white-box attack, often generates AEs with poor transferability on surrogate models. This also explains why many baseline methods relying on FGSM-generated AEs fail to effectively attack the target black-box model. In contrast, even when using FGSM to generate AEs, our method still delivers strong black-box attack performance.

4.3. Further Analysis

To further explore the effects of different components in our KOEnsAttack method, we conduct extensive ablation

Victim Dataset	Target Model	Method	Query↓	Targeted			Untargeted		
				FGSM↑	PGD↑	BIM↑	FGSM↑	PGD↑	BIM↑
SVHN	ResNet-18	DFME [30]	2M	39.62	88.62	89.81	76.68	95.86	96.18
		DFMS [25]		40.26	89.72	90.66	85.57	96.37	96.57
		DFTA [37]		8.694	17.26	23.15	33.51	41.60	45.43
		IDEAL [38]		7.286	14.40	16.92	30.81	33.83	36.17
		DisGUIDE [23]		36.24	75.15	76.90	70.93	89.46	90.13
		STDatav2 [27]		34.68	86.10	87.26	82.18	94.66	94.77
		KOEnsAttack (Ours)	**2M**	**44.59**	**92.07**	**92.90**	**87.38**	**97.41**	**97.53**
CIFAR-10	ResNet-34	DFME [30]	8M	17.63	85.24	89.10	68.42	97.14	97.80
		DFMS [25]		11.59	43.21	50.20	46.86	78.12	79.50
		DFTA [37]		8.830	27.22	30.71	37.63	52.90	56.34
		IDEAL [38]		10.08	31.81	35.79	45.41	64.12	67.63
		DisGUIDE [23]		21.81	95.32	97.11	76.94	99.60	99.83
		STDatav2 [27]		21.90	91.48	94.33	74.87	99.14	99.33
		KOEnsAttack (Ours)	**3M**	**27.79**	**97.82**	**98.66**	**82.78**	**99.78**	**99.88**
CIFAR-100	ResNet-34	DFME [30]	10M	2.278	42.12	52.71	77.79	95.21	95.40
		DFMS [25]		1.442	8.413	10.84	71.32	72.58	72.81
		DFTA [37]		1.332	11.81	16.12	62.83	73.34	74.92
		IDEAL [38]		1.610	9.264	12.63	61.10	71.63	73.95
		DisGUIDE [23]		2.581	57.20	68.24	81.81	95.98	97.10
		STDatav2 [27]		3.253	43.37	49.01	80.73	93.42	93.16
		KOEnsAttack (Ours)	**4M**	**5.021**	**60.48**	**69.79**	**85.44**	**97.85**	**98.12**
Tiny ImageNet	ResNet-50	DFME [30]	10M	0.462	2.601	3.037	49.59	57.58	61.53
		DFMS [25]		0.342	0.292	0.472	41.92	34.27	37.94
		DFTA [37]		0.000	0.020	0.000	0.000	8.479	0.000
		IDEAL [38]		0.302	0.151	0.201	41.50	24.67	32.11
		DisGUIDE [23]		0.854	1.025	2.051	62.77	39.35	49.17
		STDatav2 [27]		0.261	0.121	0.171	42.76	17.64	26.72
		KOEnsAttack (Ours)	**4M**	**2.201**	**15.22**	**23.12**	**79.34**	**78.23**	**85.65**

Table 2. Comparing ASRs results among our method and competitors with various white-box adversarial example generation methods across four datasets. For a fair comparison, we utilize the pair of VGG-13 and Inception-v3 as substitute model for all substitute training.

studies to validate the effects of key components and hyper-parameters in our method.

Ablation Studies. First, we define the following ablation terms in our experiments: (1) Baseline: maximizing the information entropy to optimize the generator and employing the cross-entropy loss to constrain the outputs' discrepancy between each student and victim model; (2) KOM: minimizing the mimic loss between each student and victim model while striving to orthogonalize the non-target class vectors from multiple students; (3) SHE: iteratively transforming the original synthetic samples through reversing the gradient for improving training efficiency of the substitute; (4) KOEnsAttack: simultaneously employing the sample hardness enhancement strategy and the knowledge orthogonalization module.

The results among the variants in Table 3 can be summarized as the following: (1) Comparing the results between the Baseline and KOM, it is evident that the knowledge orthogonalization module can help two students to better learn complementary and valuable information from black-box during model training, thereby generating more effective adversarial samples to attack the victim model. (2) With the SHE strategy, the ASRs have been significantly improved. Such results demonstrates that the sample hardness enhancement can efficiently explore the sample space under specified constraints, identifying high query-value synthetic samples after just a few iterations, which significantly re-

Type	Method	SVHN	CIFAR-10	CIFAR-100	Tiny
Targeted	Baseline	79.16	49.00	29.82	0.663
	+ KOM	85.72	55.44	36.60	1.306
	+ SHE	90.80	93.74	58.31	8.186
	Ours	**92.07**	**95.44**	**60.48**	**15.22**
Untargeted	Baseline	92.07	83.63	87.06	40.04
	+ KOM	94.94	89.37	89.40	47.65
	+ SHE	96.89	99.85	97.06	64.54
	Ours	**97.41**	**99.89**	**97.85**	**78.23**

Table 3. ASR results on variants of the proposed KOEnsAttack method. The victim models are ResNet-18 for SVHN, VGG-19 for CIFAR-10, ResNet-34 for CIFAR-100, and ResNet-50 for Tiny ImageNet, respectively.

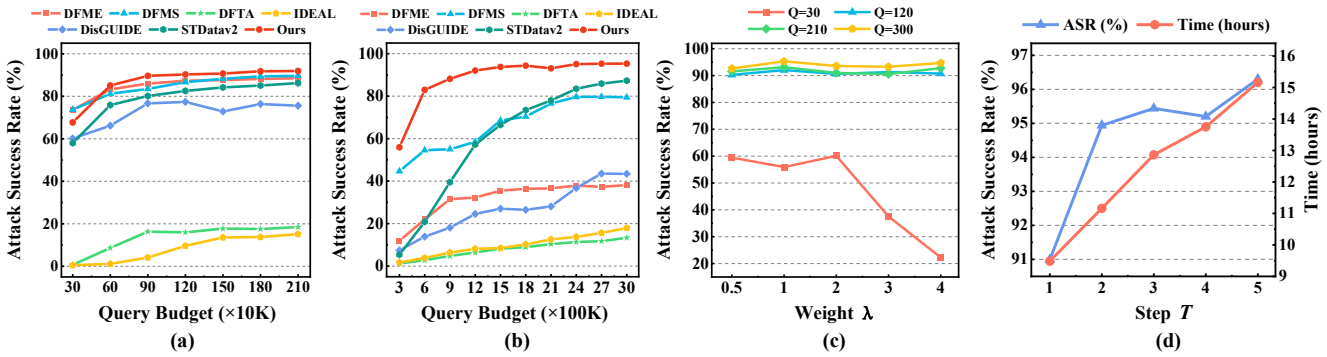

Figure 3. **Left:** Comparison of our method and baselines under different query budgets on (a) SVHN and (b) CIFAR-10. **Right:** Parameter analysis on (c) weight value λ in KOM and (d) step T in SHE on CIFAR-10. The victim models on SVHN and CIFAR-10 are ResNet-18 and VGG-19, respectively.

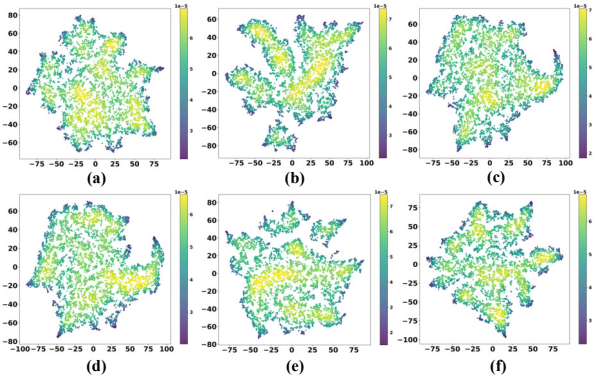

Figure 4. t-SNE visualization of synthetic data generated by (a) STDatav2, (b) DisGUIDE, (c) IDEAL, (d) DFTA, (e) DFMS, and (f) Ours.

duce black-box query budget required during model training. (3) Comparing the results between the Baseline and KOEnsAttack, it is obvious that our KOEnsAttack method achieves the best attack performance with the help of the KOM and SHE. This further validates the compatibility and effectiveness of both components.

About Hyper-Parameter Sensitivity. In the KOM, λ is the hyper-parameter utilized to control the weight corresponding to $\mathcal{L}_{ol}$. As shown in Figure 3, during the early training phase, the attack performance of adversarial samples generated on the substitute varies significantly due to different λ values. Fortunately, as the number of black-box queries increases, the black-box attack performance improves substantially across different weight λ settings, and the results gradually converge. This result also demonstrates that KOM is robust and insensitive to this hyper-parameter. In the SHE strategy, step T determines the number of iterations for sample transforming. As shown in Figure 3, the black-box attack performance steadily improves with the increasing in the number of iterations. It is important to note that more iterations also mean increased time spent on transforming samples. Therefore, in our experiments, we typically choose an appropriate step T to ensure both strong black-box attack performance and avoid excessive transforming time that could impact subsequent model training.

Visualization of Synthesized Samples. We also provide the visualization of the synthesized data for comparison in ten classes, *i.e.*, the victim model is trained on CIFAR-10 dataset. As shown in Figure 4, compared to the baseline methods, the synthesized data of our method are evenly distributed, which suggests a better sample generation for the substitute model training.

5. Conclusion

In this paper, we design a novel Knowledge-Orthogonalized Ensemble Attack (KOEnsAttack) method to tackle the data-free black-box adversarial attack task. Through the sample hardness enhancement, we can efficiently explore the data space based on original synthetic samples to obtain new high query-valuable samples, greatly improving training efficiency of substitutes. Meanwhile, the knowledge orthogonalization module introduces an effective optimization objective for students, helping each student to learn complementary and useful information, thereby training a source model more suitable for black-box adversarial attacks. The experiments over four general image classification datasets show that our KOEnsAttack not only significantly improves the attack performance against target models but also greatly reduces the required black-box query budgets.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China under Grant Nos. 62202104, 62302468, and U2441239; and the Scientific Research and Development Foundation of Fujian University of Technology, China (No. GYZ220209).

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1

[2] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4489–4498, 2023. 1, 2

[3] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2, 5

[6] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[8] Jiawei Du, Hu Zhang, Joey Tianyi Zhou, Yi Yang, and Jiashi Feng. Query-efficient meta attack to deep neural networks. In *International Conference on Learning Representations*, 2020. 2

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Nathan Inkawhich, Kevin Liang, Lawrence Carin, and Yiran Chen. Transferable perturbations of deep feature distributions. In *International Conference on Learning Representations*, 2020. 2

[12] Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13814–13823, 2021. 3, 5

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Tech Report*, 2009. 5

[14] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 5

[15] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020. 1

[16] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1

[17] Chen Ma, Li Chen, and Jun-Hai Yong. Simulating unknown target models for query-efficient black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11835–11844, 2021. 2

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 5

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1

[20] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *International Conference on Learning Representations*, 2022. 2

[21] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 5

[22] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[23] Jonathan Rosenthal, Eric Enouen, Hung Viet Pham, and Lin Tan. Disguide: Disagreement-guided data-free model extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9614–9622, 2023. 2, 3, 5, 6, 7

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[25] Sunandini Sanyal, Sravanti Addepalli, and R Venkatesh Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15284–15293, 2022. 2, 3, 5, 6, 7

[26] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[27] Xuxiang Sun, Gong Cheng, Hongda Li, Chunbo Lang, and Junwei Han. Stdatav2: Accessing efficient black-box stealing for adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3, 5, 6, 7

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1, 2

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[30] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4771–4780, 2021. 2, 3, 5, 6, 7

[31] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International conference on machine learning*, pages 5025–5034. PMLR, 2018. 2

[32] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 1

[33] Jia-Li Yin, Bin Chen, Wanqing Zhu, Bo-Hao Chen, and Ximeng Liu. Push stricter to decide better: A class-conditional feature adaptive framework for improving adversarial robustness. *IEEE Transactions on Information Forensics and Security*, 18:2119–2131, 2023. 1

[34] Jia-Li Yin, Menghao Chen, Jin Han, Bo-Hao Chen, and Ximeng Liu. Adversarial example quality assessment: A large-scale dataset and strong baseline. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 4786–4794. Association for Computing Machinery, 2024.

[35] Jia-Li Yin, Weijian Wang, Lyhwa , Wei Lin, and Ximeng Liu. Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, (9):9508–9516, 2025. 1

[36] Zheng Yuan, Jie Zhang, and Shiguang Shan. Adaptive image transformations for transfer-based adversarial attack. In *European Conference on Computer Vision*, pages 1–17. Springer, 2022. 2

[37] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022. 5, 6, 7

[38] Jie Zhang, Chen Chen, and Lingjuan Lyu. IDEAL: Query-efficient data-free learning from black-box models. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5, 6, 7

[39] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1039–1048, 2020. 2

[40] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2020. 2, 3

[41] Juntang Zhuang, Junlin Yang, Lin Gu, and Nicha Dvornek. Shelfnet for fast semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 1