Inter-Class Topology Alignment for Efficient Black-Box Substitute Attacks

Lingzhuang Meng[1], Mingwen Shao[1⋆], Yuanjian Qiao[1], and Wenjie Liu[1]

School of Computer Science and Technology, China University of Petroleum (East China), Qingdao, 266580, China
lzhmeng1688@163.com, smw278@126.com, yjqiao@s.upc.edu.cn, liuwenjie_82@163.com

Abstract. In black-box attacks based on substitute training, the similarity of the substitute model to the target model is critical for successful attacks. However, existing schemes merely train the substitute model to mimic the outputs of the target model without fully simulating the decision space, resulting in the adversarial samples generated by the substitute model being classified into the non-target class by the target model. To alleviate this issue, we propose a novel **Inter-Class Topology Alignment (ICTA)** scheme to more comprehensively simulate the target model by aligning the inter-class positional relationships of two models in the decision space. Specifically, we first design the Position Exploration Sample (PES) to more thoroughly explore the relative positional relationships between classes in the decision space of the target model. Subsequently, we align the inter-class topology between the two models by utilizing the PES to constrain the inter-class relative position of the substitute model in different directions. In this way, the substitute model is more consistent with the target model in the decision space, so that the generated adversarial samples will be more successful in misleading the target model to classify them into the target class. The experimental results demonstrate that our ICTA significantly improves attack success rate in various scenarios compared to existing substitute training methods, particularly performing efficiently in target attacks.

Keywords: Black-box attacks · Substitute attacks · Position exploration sample · Inter-class topology alignment

1 Introduction

Convolutional Neural Networks (CNNs) have shown impressive performance in various fields and are widely applied in daily life. However, it has been found that CNNs exhibit poor robustness to well-designed perturbations [3, 36, 38], which poses a great threat to their safety and reliability. For example, in the field of autonomous driving, if the network makes a wrong judgment, it will lead to potential safety hazards [13, 14, 26]. Therefore, it is especially urgent to

⋆ Corresponding author

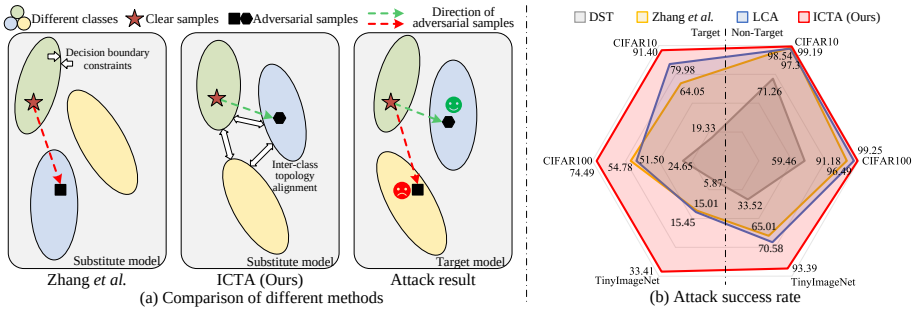


Fig. 1: (a) Comparison of different methods and **(b)** Attack success rate. Traditional approaches only focus on intra-class consistency, resulting in the adversarial samples generated by the substitute model failing to mislead the target model to classify them into the target class. In contrast, our ICTA better simulates the target model by *aligning the inter-class topology in the decision space*, resulting in a successful target attack. The performance of our ICTA is significantly improved across multiple datasets, particularly in target attacks.

research deeply into the adversarial attacks [4,30,38] and defense [17,22] schemes to enhance the reliability and security of CNNs in practical applications.

Existing adversarial attacks can be categorized into white-box and black-box attacks according to the knowledge of the attacker about the target model. In white-box attacks [3–5,11,12,16,29], the attacker possesses complete information about the target model (such as network structure, gradient, *etc.*), thus can generate highly deceptive adversarial samples based on this information. By contrast, in black-box attacks [2, 9, 21, 24, 31, 35], the attacker can only utilize the outputs of the network (probabilities or labels) to craft adversarial samples, making the task more challenging and realistic. In such scenario, the substitute attacks [9, 27, 33, 35, 37] is proposed, which focus on training a substitute model to emulate the target model, so that the adversarial samples generated by the substitute model can effectively attack the target model.

Current substitute attack methods [9, 21, 27, 28, 33, 35, 37] adhere to the paradigm of knowledge distillation, encouraging the substitute model to mimic the outputs of the target model. Some schemes utilized simple L_2 loss [21,28,33, 37] or KL-divergence [9] to constrain the substitute model. As an improvement, DST [27] leveraged a graph structure consisting of multiple outputs to optimize the substitute model. Nevertheless, these strategies only focus on learning the simplistic outputs of the target model, without adequately considering the decision space of the network. Subsequently, Zhang *et al*. [35] considered aligning the decision boundary between the substitute model and the target model to bridge the gap between the two models. However, this restriction of decision boundary only aims at the intra-class decision consistency and *ignores the inter-class relationship consistency of the two models in the decision space*. As a result, the adversarial samples generated by the substitute model are difficult to be classified as target class by the target model, as Fig. 1(a).

To alleviate the above issue, we propose a novel **Inter-Class Topology Alignment (ICTA)** scheme to align the inter-class positional relationships of the two models in decision space, thus better simulating the target model and achieving more accurate target attacks. Specifically, we first design the Position Exploration Sample (PES) to explore the relative positional relationships among all classes of the target model by introducing adversarial perturbations of different directions and strengths. Subsequently, we constrain the inter-class positional consistency of the substitute model to align the inter-class topology between the two models. Thanks to the better alignment of the two models in the decision space, the adversarial samples generated by the substitute model can be more successful deceiving the target model to classify them into the target class, as shown in Fig. 1(a). Experiments demonstrate that our ICTA achieves higher attack success rate on different datasets, and especially more pronounced in target attacks, as Fig. 1(b). Our primary contributions are as follows:

- We propose ICTA to innovatively align the substitute model with the target model in terms of inter-class topology in the decision space, which effectively improves the success rate of black-box substitute attack.
- We propose a novel PES to more comprehensively explore the inter-class relative position of the target model, which is utilized to align the inter-class topology between the substitute model and the target model.
- Experiments illustrate that our ICTA achieves significant performance gains in a variety of scenarios and outperforms on target attacks compared to the existing substitute attack schemes.

2 Related Works

2.1 White-box attacks

In white-box attacks, the attacker is fully aware of the information about the target model, such as the structure and parameters [5, 11, 12, 16], and highly deceptive adversarial samples can be generated based on this information. Classical white-box adversarial sample generation schemes utilized gradient information of network to generate adversarial samples, such as FGSM [5], BIM [11], and PGD [16]. Moreover, Croce *et al.* [4] proposed parameter-free attack combination AutoAttack by integrating multiple attack methods for testing adversarial robustness. However, white-box attacks may lack practicality in real-world scenarios due to the requirement for extensive knowledge about the target model.

2.2 Black-box attacks

In contrast, black-box attacks are more challenging and more in line with realistic scenarios, as the attacker can generate adversarial samples only using the outputs of the target model, including transfer attacks, query attacks, and substitute attacks. Transfer attacks [1, 3, 7, 29, 31] aim to maximize the transferability of existing adversarial samples, enabling them to successfully attack the target

model. For example, Zhu *et al.* [38] proposed modifying the distribution of the original images to match the distribution of the target class, thereby obtaining adversarial samples with high transferability. Query attacks [2, 24, 30] obtain adversarial samples by continuously querying the target model using specific inputs until a desired output is obtained. Substitute attacks [21, 35, 37] focus on training a substitute model that is similar to the target model, so that the adversarial samples generated by the substitute model can successfully attack the target model. Compared to other schemes, the substitute attack is able to generate more effective adversarial samples due to the ability to simulate the decision process of the target model.

Most of the existing substitute attack schemes used L_2 loss [21, 28, 33, 37] or KL-divergence [9] as the constraint to train substitute models. For example, DaST [37] used Generative Adversarial Networks (GANs) to generate data for substitute training and leveraged the L_2 loss as a constraint to align the outputs of the substitute model with the target model. Kariyappa *et al.* [9] optimized the generator with the help of zeroth-order gradient estimate and trained the substitute training model with a simple KL-divergence. Shao *et al.* [21] proposed LCA to utilize a diffusion model to generate the data required for substitute training, which greatly improved the training efficiency compared to traditional GANs-based methods. However, the target attack success rate is not impressive as it only used L_2 loss to constrain the substitute model. In order to simulate the target model more precisely, Wang *et al.* [27] proposed DST to optimize the substitute model from a graph with multiple outputs of the target model instead of a single output. Nevertheless, these schemes only learn the simple outputs of the target model without fully considering the complex decision space attributes, resulting in poor performance. Subsequently, Zhang *et al.* [35] encouraged a high alignment of decision boundary between the substitute model and the target model, which improved the attack success rate.

However, the aforementioned schemes only focused on the intra-class consistency and neglected the inter-class relational consistency, resulting in the adversarial samples generated by the substitute model being difficult to be classified as target class by the target model. In this paper, we further consider the inter-class positional relationships in the decision space and better simulate the target model by aligning the inter-class topology for a more effective target attack.

3 Methodology

3.1 Preliminary

For a black-box target model T , existing research on substitute attacks typically trains a substitute model S similar to the T through a knowledge distillation paradigm. Formally, the substitute model can be obtained by substitute training:

$$\arg\min_{\theta} \mathbb{E}_{x\sim\mathcal{X}} d[S(x), T(x)], \tag{1}$$

where d denotes the distance metric function, $x \sim \mathcal{X}$ denotes the training data, and θ is the parameters of S. The goal of substitute training is to optimize the substitute model to resemble the target model as similar as possible.

Subsequently, the adversarial samples can be generated using white-box adversarial sample generation methods based on the substitute model. We take BIM [11] as an instance, which perturbs the input data along the gradient direction of the target class through multiple iterations to make the output of the model close to the target class in the decision space, as Fig. 2. The generation of adversarial samples for target attacks can be represented as follows:

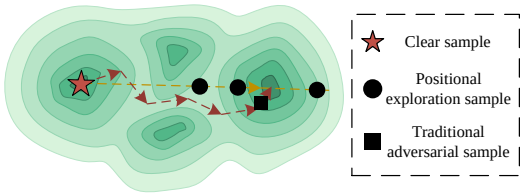


Fig. 2: Position exploration sample vs. Traditional adversarial sample. Traditional adversarial samples are closer to and clustered at the decision center after several iterations of minor adjustments. While our PES has a much broader coverage and is able to explore the decision space more comprehensively.

$$x_0^{adv} = x,\ x_{N+1}^{adv} = Clip_{x,\epsilon}\{x_N^{adv} - \alpha sign(\nabla_x L(S(x_{N+1}^{adv}), y_{tar}))\}, \tag{2}$$

where y_{tar} is the target class, ∇_x denotes the gradient of the substitute model, $sign$ denotes the sign function, α represents the step size during the iteration process, and $Clip$ is the clip function based on the perturbation boundary ϵ.

Finally, the generated adversarial samples are used to attack the target model. Clearly, the similarity between the substitute model and the target model is a key factor for successful attacks in substitute attacks.

3.2 Aligning the Inter-Class Topology in the Decision Space

Existing substitute training methods encourage the substitute model to simply mimic the outputs of the target model without fully simulating the decision space of the target model. Resulting in the adversarial samples generated by the substitute model are difficult to be classified into the target class by the target model. Moreover, there is also a lack of research on what aspects the substitute model needs to resemble the target model. In this paper, we propose to understand and improve the substitute training from the perspective of the inter-class topology in the decision space, which refers to the positional relationships (including directions and distances) among all classes in the decision space.

According to Eq. (2), modifying the adversarial samples along the direction of the gradient ∇_x in the decision space can bring the decision of the model closer to the target class, as shown in Fig. 2. Therefore, if the relative positions of all classes are consistent between the substitute model and the target model, then the adversarial samples generated by the substitute model can be successfully classified into the target class by the target model. This suggests that aligning

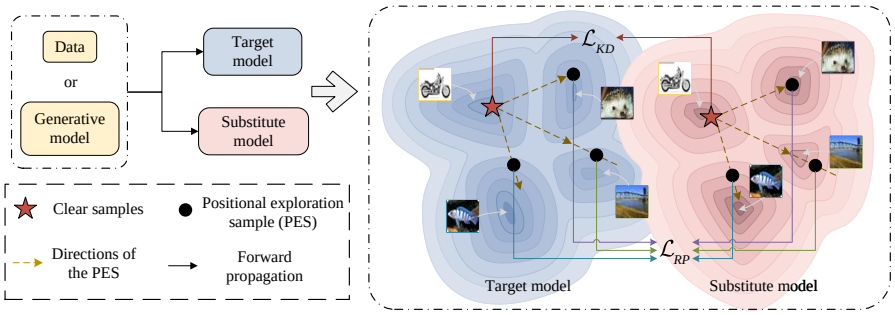


Fig. 3: The framework of our **Inter-Class Topology Alignment (ICTA)**. ICTA explores the inter-class relative position in the decision space using position exploration sample in different directions, and aligns the inter-class topology of the two models by fitting the inter-class relative position.

the inter-class topology between the two models in the decision space is crucial for substitute attacks.

For the sample x_i, we consider $S(x_i)$ and $T(x_i)$ as nodes in the two decision spaces, and $e_{S_{ij}} = \langle S(x_i), S(x_j) \rangle, i \neq j$ and $e_{T_{ij}} = \langle T(x_i), T(x_j) \rangle, i \neq j$ to represent the edges of the topology in the two decision spaces. Topology consistency can be achieved by aligning all nodes and edges:

$$\min \sum_{i=1}^{n} \mathcal{D}_p(S(x_i), T(x_i)) + \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{D}_e(e_{S_{ij}}, e_{T_{ij}}), \tag{3}$$

where $\mathcal{D}_p$ and $\mathcal{D}_e$ denote the distance metrics of points and edges, respectively.

However, aligning the edges among all the samples in the decision space is extremely complex. Therefore, we propose a relative position exploration and alignment strategy for matching the inter-class topology of the two models.

3.3 Position Exploration Sample

According to Fig. 3, the relative position of the clean and adversarial samples in the decision space can similarly be used to represent the relative position of the two classes in the decision space. In this case, the adversarial perturbations used to generate the adversarial samples can reflect the distance and direction between the two classes. Therefore, under the assumption that the target model is white-boxed, the inter-class relative position can be estimated using the positional relationships between the clean and adversarial samples.

The traditional adversarial samples aim to progressively bring the sample closer to the decision center of the target class through multiple iterations of minor adjustments. While the sample has a relatively high success rate in the attack, it does not beneficial for thoroughly exploring the entire decision space. To alleviate this limitation, we design the Position Exploration Sample (PES)

to more comprehensively explore the inter-class relative position of the target model in the decision space, which can be represented as:

$$x_{pes} = x - \phi \nabla_x L(T(x), y_{tar}), \tag{4}$$

where ϕ is the perturbation strength. We generate the PES using the full gradient, not just the gradient direction, without iteration and without being constrained by the pixel range. Thus, the gradient ∇_x and the perturbation strength ϕ determine its direction and distance relative to the clean sample in the decision space. We represent the inter-class relative position based on the relative position of the PES to the clean sample and explore the decision space more thoroughly over a wider coverage by controlling y_{tar} and ϕ, as Fig. 2.

However, it is impossible to generate the PES directly due to the gradient of the target model is inaccessible. Therefore, we retreat to generating PES using the substitute model, as shown in Fig. 3, which can be represented as:

$$x_{pes} = x - \phi \nabla_x L(S(x), y_{arb}), \tag{5}$$

where y_{arb} represents an arbitrary class different from ground truth. In this way, we can explore the decision space from different directions and obtain the complete inter-class topology in the decision space. Moreover, the PES generated by the substitute model has an effect on exploring the decision space of the target model, since our goal is to make the two models as similar as possible.

3.4 Topology Alignment Based on PES

In order to align the inter-class topology of the two models, we simplify the topology consistency in Eq. (3) to the consistency of the inter-class relative position. This method allows us to capture the inter-class topology through the position of two nodes without the need to optimize the edges directly, as demonstrated in Fig. 3. We can achieve this alignment by simultaneously optimizing the knowledge distillation loss L_{KD} and the relative position loss L_{RP}:

$$L_{KD} = \mathcal{D}_{kd}(T(x), S(x)), \tag{6}$$

$$L_{RP} = \mathcal{D}_{rp}(T(x_{pes}), S(x_{pes})), \tag{7}$$

$$L_T = \lambda_1 L_{KD} + \lambda_2 L_{RP}, \tag{8}$$

where $\mathcal{D}_{kd}$ and $\mathcal{D}_{rp}$ are distance measure formulas, λ_1 and λ_2 are hyperparameters. The inter-class relations of the two models can be aligned by ensuring intra-class consistency through L_{KD} and inter-class consistency through L_{RP}.

Furthermore, we can constrain the substitute model to be inter-class consistent with the target model in different directions and distances based on the PES of a clean sample in different directions and at different distances. At this point, the loss of topology consistency can be expressed as:

$$L_T = \lambda_1 L_{KD} + \lambda_2 \sum_{i=1}^{M} L_{RP}^i, \tag{9}$$

Table 1: The accuracy of target model. We selecte four models with different architectures as target models on each of the three datasets.

Dataset	CIFAR10			
Target Model	ResNet34	Vgg16	GoogleNet	ShuffleNetV2
Accuracy	94.53	92.82	95.31	94.53
Dataset	CIFAR100			
Target Model	ResNet50	MobileNetV2	WideResNet	ResNeXt50
Accuracy	79.3	71.61	79.69	77.56
Dataset	TinyImageNet			
Target Model	ResNet50	Vgg19	MobileNetV2	ResNeXt50
Accuracy	62.81	59.66	62.81	66.33

where M is the number of x_{pes} generated for a same clean sample.

In addition, incorporating the original categorical loss into this process is often beneficial, which is usually the cross-entropy loss L_{CE}. The overall loss can be expressed as:

$$L_{all} = L_T + \lambda_3 L_{CE}, \tag{10}$$

where λ_3 is a hyperparameter. The parameters of the substitute model are optimized by back propagation to better simulate the target model.

4 Experimental

4.1 Experimental Details

Dataset and Target Model. We choose CIFAR10 [10] (with size of 32×32 and 10 classes), CIFAR100 [10] (with size of 32×32 and 100 classes), TinyImageNet [19] (with size of 64×64 and 200 classes) as the training dataset. And we select VGG(16, 19) [23], ResNet(34, 50) [6], MobileNetV2 [20], GoogleNet [25], WideResNet [34], ShuffleNetV2 [15], and ResNeXt50 [32] as the target model. The accuracy of the target models is shown in Tab. 1.

Training Setup. During training, we set the perturbation strength in Eq. (5) to $\phi = |Y|$, where Y obeys the Gaussian distribution $Y \sim N(0, 3)$. The number of PES in Eq. (9) is set to $M = 1$. The hyperparameters are empirically set to $\lambda_1 = \lambda_3 = 1, \lambda_2 = 10$. We standardize the query budget for substitute training across all baselines to ensure fairness, setting it to 2 million for CIFAR10 and 3 million for CIFAR100 and TinyImageNet.

Additionally, we conduct experiments in two scenarios: full-data and data-free. In the data-free scenario, we utilize Stable Diffusion [18] as a generator and use the LCA [21] strategy to produce the data needed for substitute training. In the full-data scenario, the whole training data is available for substitute training. We perform this part of the experiment to validate the performance of the scheme without the influence of the generator. The algorithm of our ICTA and other experimental details are described in the **supplement**. The source code will be released upon acceptance of the paper.

Table 2: Comparison of different white-box adversarial sample generation schemes. We choose ResNet34 on CIFAR10, ResNet50 on CIFAR100 and VGG19 in TinyImageNet as target models. The best results are marked as **bolded**.

		Dataset	CIFAR10			CIFAR100			TinyImageNet		
		Methods	FGSM	BIM	PGD	FGSM	BIM	PGD	FGSM	BIM	PGD
Non-Target	GANs	MAZE [9]	38.64	57.98	68.94	38.20	41.55	43.07	9.18	9.79	11.10
		DaST [37]	35.19	56.21	64.22	40.01	42.54	47.40	9.71	9.26	9.81
		DST [27]	40.83	67.65	71.26	46.45	49.21	59.46	11.67	12.47	33.52
		Zhang *et al.* [35]	60.11	94.21	97.54	83.69	94.53	91.18	38.58	48.21	65.01
	Diffusion	LCA [21]	68.29	95.73	97.30	82.03	95.58	96.49	51.55	52.35	70.58
		ICTA (Ours)	**75.60**	**98.71**	**99.19**	**87.42**	**99.01**	**99.25**	**83.74**	**92.73**	**93.39**
Target	GANs	MAZE [9]	6.84	13.83	16.65	6.20	12.78	10.01	0.50	2.50	2.54
		DaST [37]	7.85	15.41	18.61	5.01	12.48	11.41	0.20	3.02	5.20
		DST [27]	10.26	16.12	19.33	**10.22**	16.37	24.65	1.88	3.05	5.87
		Zhang *et al.* [35]	28.17	72.05	64.05	7.74	58.45	54.78	2.30	10.88	15.01
	Diffusion	LCA [21]	24.31	76.90	79.98	4.61	45.19	51.50	1.84	12.95	15.45
		ICTA (Ours)	**29.41**	**87.63**	**91.40**	7.80	**69.02**	**74.49**	**6.87**	**29.31**	**33.41**

Evaluation Setup and Process. During testing, we choose three classical methods, including FGSM [5], BIM [11], and PGD [16], to generate adversarial samples, with parameters uniformly set to perturbation boundary $\epsilon = 8/255$ and step size $\alpha = 2/255$. Our hardware utilizes an RTX 2080Ti GPU, and all schemes are retrained under the same environment.

We train substitute model by our ICTA and then generate adversarial samples based on the substitute models using a white-box adversarial sample generation scheme. The success rate of these adversarial samples in attacking the target model is used as the final evaluation metric.

4.2 Comparative Experiments in Data-free Scenario

Comparison of White-box Adversarial Sample Generation Schemes. We compare different white-box adversarial sample generation schemes, including FGSM, BIM, and PGD. As illustrated in Tab. 2, our ICTA realizes significant performance improvements across different white-box adversarial sample generation methods. On the one hand, our ICTA achieves more noticeable performance enhancement on large datasets TinyImageNet than on small datasets CIFAR10. On the other hand, ICTA has superior performance improvement on target attacks than non-target attacks. The results indicate that our ICTA more effectively simulates the decision space of the target model compared to traditional substitute training methods [9, 21, 27, 35, 37].

Comparison on Different Target Models. According to the results in Tab. 3, our ICTA demonstrates a significant performance improvement on different target models compared to the GANs-based approaches [9, 27, 35, 37] and the diffusion-based method [21]. Specifically, ICTA achieves the most superior performance in target attacks, especially on TinyImageNet. Additionally, ICTA

Table 3: Comparison on different target models. We choose PGD as the default white-box adversarial sample generation scheme. The best results are marked as **bolded**.

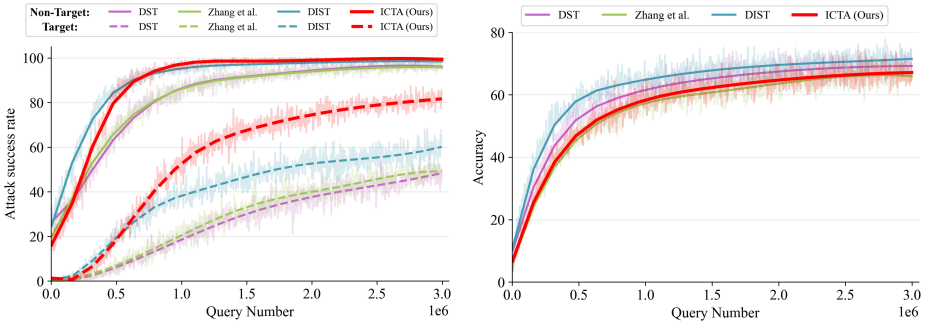
		Dataset	CIFAR10			CIFAR100		TinyImageNet	
		Target Model	VGG16	ResNet34	GoogleNet	MobileNet	ResNet50	ResNet50	VGG19
Non-Target	GANs	MAZE [9]	47.49	68.94	27.13	51.73	43.07	15.54	11.10
		DaST [37]	53.29	64.22	26.48	57.94	47.40	16.11	9.81
		DST [27]	55.15	71.26	38.64	65.85	59.46	32.51	33.52
		Zhang *et al.* [35]	95.84	97.54	67.24	97.78	91.18	78.82	65.01
	Diffusion	LCA [21]	**97.01**	97.30	94.82	97.01	96.49	67.56	70.58
		ICTA (Ours)	96.48	**99.19**	**97.32**	**99.63**	**99.25**	**94.80**	**93.39**
Target	GANs	MAZE [9]	11.39	16.65	12.98	11.03	10.01	3.18	2.54
		DaST [37]	15.39	18.61	5.31	9.58	11.41	2.21	5.20
		DST [27]	17.07	19.33	18.89	18.46	24.65	8.55	5.87
		Zhang *et al.* [35]	57.47	64.05	39.83	44.11	54.78	20.01	15.01
	Diffusion	LCA [21]	75.39	79.98	78.18	51.61	51.50	12.45	15.45
		ICTA (Ours)	**79.57**	**91.40**	**84.57**	**70.56**	**74.49**	**43.02**	**33.41**

Table 4: Comparison of attack success rate in label-based and probability-based scenarios. The best results are marked as **bolded**.

		method	Probability-based		Label-based	
			cifar10	cifar100	cifar10	cifar100
Non-Target	GANs	DaST [37]	64.22	47.40	67.72	36.57
		DST [27]	71.26	59.46	80.42	48.50
		Zhang *et al.* [35]	97.54	91.18	98.98	90.20
	Diffusion	LCA [21]	97.30	96.49	98.48	94.55
		ICTA (Ours)	**99.19**	**99.25**	**99.49**	**98.53**
Target	GANs	DaST [37]	18.61	11.41	19.28	5.20
		DST [27]	19.33	24.65	23.85	4.59
		Zhang *et al.* [35]	64.05	54.78	69.25	40.17
	Diffusion	LCA [21]	79.98	51.50	82.18	49.85
		ICTA (Ours)	**91.40**	**74.49**	**93.10**	**61.71**

consistently enhances performance across different target models, showcasing the stability of the performance in various scenarios. In summary, our ICTA is able to simulate different target models in black-box attack scenarios and exhibits substantial performance improvements on large-scale datasets.

Comparison in Probability-based and Label-based Scenarios. We compare the attack success rate in the probability-based and label-based scenarios, as displayed in Tab. 4. In the label-based scenario, only the output labels of the target model are accessible, while in the probability-based scenario, both the labels and probabilities of the target model are available. The experiment results indicate that our ICTA achieves the optimal performance in both probability-based and label-based scenarios and shows significant improvement over the existing substitute attack methods [21, 27, 35, 37]. This highlights the comprehensive superiority of our ICTA, demonstrating outstanding attack performance even in scenarios where only label outputs are accessible.



(a) Attack success rate **(b)** Accuracy of substitute model

Fig. 4: Comparison of attack success rate and accuracy in full-data scenario. Comparison schemes include the substitute training schemes DST [27] and Zhang *et al.* [35] as well as a distillation method DIST [8]. The ResNet50 on CIFAR100 is used as the target model.

4.3 Comparative Experiments in Full-data Scenario

Performance of Different Substitute Training Strategies. In Fig. 4, we present the attack success rate (both target and non-target attacks) and the accuracy of the substitute training during the training process. It can be observed that our ICTA demonstrates excellent performance in both target and non-target attacks and shows higher efficiency in target attacks. There is a significant improvement in non-target attacks compared to conventional methods DST [27] and Zhang *et al.* [35], which highlights the unique advantages of our ICTA in improving the effectiveness of target attacks. In addition, we compare an excellent knowledge distillation scheme, DIST [8], which also shows the superior performance of our ICTA in attack scenarios.

Notably, we find that the accuracy of the substitute model does not fully reflect the effectiveness of the substitute training. Despite having a relatively low accuracy in our ICTA, the substitute model exhibits excellent performance in attacks. This indicates that our ICTA better simulates the target model in terms of inter-class topology, rather than just replicating its outputs.

Comparison of Relative Performance with White-box Attacks. We further compare the attack success rate of our ICTA with the corresponding white-box attacks, as demonstrated in Tab. 5, where the Relative Performance (RP) is used as a metric to measure the performance relative to the white-box attacks. Our ICTA achieves comparable performance to the corresponding white-box attacks in non-target attack, when the query budgets is set to 2M (for CIFAR10) and 3M (for CIFAR100 and TinyImageNet). Moreover, it is worth noting that the target attack success rate exceeds 83% on the CIFAR10 dataset, 67% on the CIFAR100 dataset, and 44% on the TinyImageNet dataset compared to the corresponding white-box attack. As can be seen, the proposed ICTA achieves quite

Table 5: Comparison with direct white-box attacks on target model. We utilize Relative Performance (RP) to evaluate the attack success of our black-box attack method relative to the white-box attack (in red).

			ResNet34			Vgg16			GoogleNet			ShuffleNetV2		
			white-box	Ours	RP	white-box	Ours	RP	white-box	Ours	RP	white-box	Ours	RP
CIFAR10	Non-Target	FGSM	78.54	83.42	1.06	85.59	79.46	0.93	63.84	77.73	1.22	84.97	83.97	0.99
		BIM	100.00	99.59	1.00	100.00	97.26	0.97	99.93	97.61	0.98	100.00	98.86	0.99
		PGD	100.00	99.79	1.00	100.00	98.08	0.98	100.00	98.26	0.98	100.00	99.20	0.99
	Target	FGSM	28.79	34.69	1.20	19.66	30.25	1.54	24.55	26.99	1.10	32.93	31.78	0.97
		BIM	99.84	94.58	0.95	84.81	81.19	0.96	99.55	82.58	0.83	100.00	86.01	0.86
		PGD	100.00	96.63	0.97	98.21	85.67	0.87	99.97	86.29	0.86	100.00	89.03	0.89
			ResNet50			WideResNet			MobileNetV2			ResNeXt50		
CIFAR100	Non-Target	FGSM	84.77	86.76	1.02	74.27	83.40	1.12	90.18	88.20	0.98	87.24	87.51	1.00
		BIM	100.00	99.22	0.99	98.11	96.83	0.99	99.98	99.09	0.99	99.98	99.16	0.99
		PGD	100.00	99.50	0.99	99.57	97.54	0.98	100.00	99.46	0.99	100.00	99.47	0.99
	Target	FGSM	3.05	8.52	2.79	3.87	9.42	2.43	4.39	7.35	1.67	3.83	8.79	2.30
		BIM	95.22	72.79	0.76	89.44	65.90	0.74	96.30	64.79	0.67	97.06	73.29	0.76
		PGD	99.25	79.34	0.80	98.38	73.22	0.74	99.66	70.33	0.71	99.70	80.15	0.80
			ResNet50			Vgg19			MobileNetV2			ResNeXt50		
TinyImageNet	Non-Target	FGSM	95.82	86.07	0.90	92.82	86.74	0.93	98.42	88.00	0.89	96.62	83.65	0.87
		BIM	99.97	96.26	0.96	99.64	95.94	0.96	100.00	97.21	0.97	99.98	94.78	0.95
		PGD	99.98	96.96	0.97	99.91	96.56	0.97	100.00	97.78	0.98	99.98	95.86	0.96
	Target	FGSM	10.25	10.04	0.98	4.79	7.99	1.67	15.71	8.74	0.56	13.18	8.68	0.66
		BIM	98.96	52.30	0.53	85.66	39.31	0.46	99.72	45.92	0.46	99.53	43.56	0.44
		PGD	99.74	57.85	0.58	97.81	46.31	0.47	99.92	51.18	0.51	99.96	50.30	0.50

satisfactory performance in black-box attack scenarios, demonstrating robust adaptability in target attacks across different datasets.

4.4 Analysis of Decision Boundary Similarity

The t-SNE Visualization of Decision Space. We exhibit the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization of the decision space of the models, as exhibited in Fig. 5. In t-SNE visualization, each data point is typically labeled with its corresponding class. Therefore, it can depict the inter-class relative position in the decision space, as well as the inter-class topology. The visualization illustrates that the inter-class topology in the decision space of the substitute model trained by our ICTA

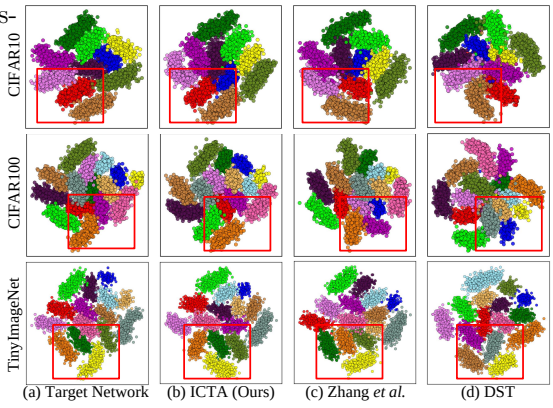
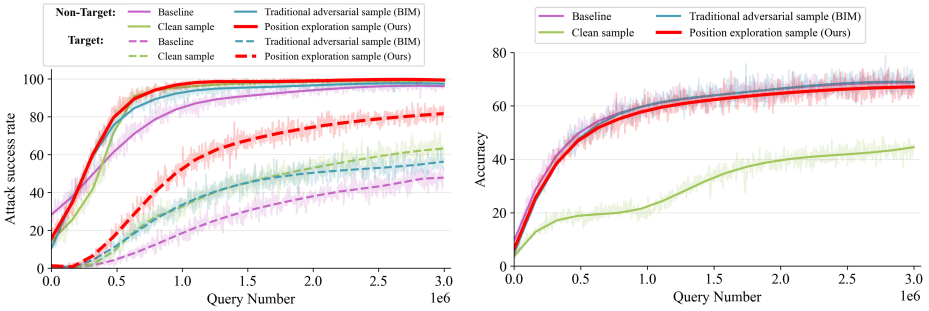


Fig. 5: The t-SNE visualization of the decision space for **(a)** the target model, **(b)** our ICTA, **(c)** Zhang *et al.* [35] and **(d)** DST [27].



(a) Attack success rate **(b)** Accuracy of substitute model

Fig. 6: Ablation experiments on PES. The experiments include baseline (no relative position loss) and strategies for exploring inter-class relative position through traditional adversarial samples, clean samples, and position exploration sample.

method is remarkably similar to the topology of the target model. On the contrary, both the output alignment-focused DST [27] and the decision boundary-constrained approach by Zhang *et al.* [35] exhibit greater differences from the target model in t-SNE visualization. This confirms that our ICTA is more effective in aligning the substitute model to match the target model.

Quantitative Analysis. We explore random directions of arbitrary samples, and then quantitatively assess the Decision Boundary Similarity (DBS) based on the overlap between target and substitute models. As in Tab. 6, our ICTA has more consistent decision boundaries with the target model.

Table 6: Quantitative analysis of decision boundary similarity.

Method	**ICTA (Ours)**	Zhang *et al.*	DST
CIFAR10	**0.93**	0.87	0.80
CIFAR100	**0.88**	0.77	0.71
TinyImageNet	**0.82**	0.57	0.55

4.5 Ablation Study

Ablation on Position Exploration Sample. In this paper, we design PES to explore the inter-class relative position in the decision space through their relative position relationships with respect to the original clean samples. However, as the position of any two samples in the decision space inherently depict the inter-class relative position, traditional adversarial sample or even a clean sample has the effect of exploring the decision space. In this section, we validate the effectiveness of PES by comparing it with traditional adversarial sample generated by BIM and clean sample, as illustrated in Fig. 6.

It is worth noting that while all three approaches contribute to explore the inter-class relative position in the decision space, PES stands out as particularly advantageous. This is due to the fact that both clean samples and the adversarial

Table 7: Ablation comparison of the loss function and the distance metric. We select two distance metrics, MSE and DIST [8], and use different combinations of distance metrics in the loss function. The best result are marked as **bolded**, and the sub-optimal results are marked as underlined.

$\mathcal{D}_{kd}$		$\mathcal{D}_{rp}$		Target			Non-Target		
MSE	DIST	MSE	DIST	FGSM	BIM	PGD	FGSM	BIM	PGD
✓				4.31	42.14	46.14	79.48	95.37	96.38
	✓			4.11	55.52	57.67	82.85	98.30	98.80
✓		✓		7.64	64.20	71.70	84.35	98.43	99.06
	✓	✓		7.82	**73.42**	79.09	86.73	99.04	**99.61**
	✓		✓	**8.52**	72.79	**79.34**	**86.76**	**99.22**	99.50

samples generated by BIM are closer to the decision center, resulting in limited exploration in the decision space. In contrast, the proposed PES is more diverse by employing the full-gradient and unconstrained setting, enabling exploration across a wider range of the decision space. The ablation experiments for the parameters of PES and the number of PES are provided in the **supplement**.

Ablation on Loss and Metrics. We conduct an ablation study on the loss functions and distance metrics of substitute training in Tab. 7, focusing on the distillation loss L_{KD} and relative position loss L_{RP}. Here, the term 'DIST' refers to the distance metric function in the distillation scheme DIST [8]. It can be seen that when DIST [8] is used as the distance metric in L_{KD}, there is a certain improvement in the attack success rate. And when L_{RP} is added, there is a significant improvement in the attack success rate, which underscores the indispensability of the relative position loss. Notably, our ICTA employs DIST as the distance metric in both losses, as it further improves attack performance.

5 Conclusion

In this paper, we propose a novel ICTA scheme to better align the substitute model to the target model for efficient black-box substitute attacks. Compared to the existing schemes that only align intra-class consistency, we simulate the target model in the decision space more comprehensively through inter-class topology alignment. Specifically, we first explore the inter-class relative position of the target model in the decision space by the designed PES. Subsequently, we achieve inter-class topology alignment of the two models by constraining the inter-class relative position of the substitute model in different directions. Benefiting from the better alignment of the two models in the decision space, the adversarial samples generated by the substitute model are more successful in misleading the target model to classify them into the target class. Experimental results demonstrate the effectiveness of ICTA, which achieves better simulation of the target model and significantly improves the attack success rate beyond existing substitute attack approaches.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFA1000102), the National Natural Science Foundation of China (Grant Nos. 62376285, 62272375, and 61673396), and Natural Science Foundation of Shandong Province, China (Grant No. ZR2022MF260).

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square Attack: a query-efficient black-box adversarial attack via random search. In: Proceedings of the European Conference on Computer Vision. pp. 484—501 (2020)
2. Brendel, W., Rauber, J., Bethge, M.: Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In: International Conference on Learning Representations. arXiv preprint arXiv:1712.04248 (2018)
3. Chen, S., He, Z., Sun, C., Yang, J., Huang, X.: Universal Adversarial Attack on Attention and the Resulting Dataset DAmageNet. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2188–2197 (2022)
4. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International Conference on Machine Learning. vol. 119, pp. 2206–2216 (2020)
5. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
7. Huang, H., Chen, Z., Chen, H., Wang, Y., Zhang, K.: T-SEA: Transfer-based Self-Ensemble Attack on Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20514–20523 (2023)
8. Huang, T., You, S., Wang, F., Qian, C., Xu, C.: Knowledge distillation from a stronger teacher. In: Advances in Neural Information Processing Systems. vol. 35, pp. 33716–33727 (2022)
9. Kariyappa, S., Prakash, A., Qureshi, M.K.: Maze: Data-free model stealing attack using zeroth-order gradient estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13814–13823 (2021)
10. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Tront (2009)
11. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
12. Lee, B.K., Kim, J.: Mitigating Adversarial Vulnerability through Causal Parameter Estimation by Adversarial Double Machine Learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4499–4509 (2023)
13. Li, L., Lian, Q., Chen, Y.C.: Adv3D: Generating 3D Adversarial Examples in Driving Scenarios with NeRF. arXiv preprint arXiv:2309.01351 (2023)
14. Liu, J., Lu, B., Xiong, M., Zhang, T., Xiong, H.: Adversarial Attack with Raindrops. arXiv preprint arXiv:2302.14267 (2023)
15. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In: Proceedings of the European Conference on Computer Vision. pp. 122–138 (2018)

16. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2018)
17. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial Attacks and Defenses in Deep Learning. Engineering **6**(3), 346–360 (2020)
18. Robin, R., Andreas, B., Dominik, L., Patrick, E., Bj\"orn, O.: High-Resolution Image Synthesis with Latent Diffusion Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10674–10685 (2022)
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
21. Shao, M., Meng, L., Qiao, Y., Zhang, L., Zuo, W.: Data-free black-box attack based on diffusion model. arXiv preprint arXiv:2307.12872 (2023)
22. Sheng, L., Liang, J., He, R., Wang, Z., Tan, T.: AdaptGuard: Defending Against Universal Attacks for Model Adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19093–19103 (2023)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Su, J., Vargas, D.V., Sakurai, K.: One Pixel Attack for Fooling Deep Neural Networks. IEEE Transactions on Evolutionary Computation **23**(5), 828–841 (2019)
25. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
26. Wang, N., Luo, Y.: Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4412–4423 (2023)
27. Wang, W., Qian, X., Fu, Y., Xue, X.: Dst: Dynamic substitute training for data-free black-box attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14361–14370 (2022)
28. Wang, W., Yin, B., Yao, T., Zhang, L., Fu, Y., Ding, S., Li, J., Huang, F., Xue, X.: Delving into data: Effectively substitute training for black-box attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4761–4770 (2021)
29. Wang, X., Zhang, Z.: Structure Invariant Transformation for better Adversarial Transferability. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4607–4619 (2023)
30. Wang, X., Zhang, Z., Tong, K., Gong, D., He, K., Li, Z., Liu, W.: Triangle Attack: A Query-efficient Decision-based Adversarial Attack. In: Proceedings of the European Conference on Computer Vision. pp. 156–174 (2022)
31. Wang, Z., Yang, H., Feng, Y., Sun, P., Guo, H., Zhang, Z., Ren, K.: Towards Transferable Targeted Adversarial Examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
32. Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K.: Aggregated Residual Transformations for Deep Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5987–5995 (2017)

33. Yu, M., Sun, S.: FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks. Computers & Security **113**, 102555 (2022)
34. Zagoruyko, S., Komodakis, N.: Wide Residual Networks. In: Procedings of the British Machine Vision Conference. pp. 87.1–87.12 (2016)
35. Zhang, J., Li, B., Xu, J., Wu, S., Ding, S., Zhang, L., Wu, C.: Towards efficient data free black-box adversarial attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15094–15104 (2022)
36. Zheng, X., Fan, Y., Wu, B., Zhang, Y., Wang, J., Pan, S.: Robust Physical-World Attacks on Face Recognition. Pattern Recognition **133**, 109009 (2023)
37. Zhou, M., Wu, J., Liu, Y., Liu, S., Zhu, C.: Dast: Data-free substitute training for adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 234–243 (2020)
38. Zhu, Y., Chen, Y., Li, X., Chen, K., He, Y., Tian, X., Zheng, B., Chen, Y., Huang, Q.: Toward Understanding and Boosting Adversarial Transferability From a Distribution Perspective. IEEE Transactions on Image Processing **31**, 6487–6501 (2022)