Supplementary Material: Inter-Class Topology Alignment for Efficient Black-Box Substitute Attack

A Additional Setup Details

During the training phase, we employ common data augmentation operations, such as Center Crop, Random Flip, Gaussian Blur *etc.*, to fully utilize of the data, and the same operations are applied in the comparison experiments for fairness. We set the iterations of PGD to 40 and used DIST as the distance metric in Eqs. (6) and (7), which was consistent across all experiments. In the label-only scenario, we used the cross-entropy function instead of DIST. The substitute model used in this paper is a simple optimized ResNet [9] architecture. We optimized the substitute model using the SGD optimizer with an initial learning rate of 0.001.

The evaluation is performed on a test set of the corresponding dataset, and we use accurately categorized images as experimental subjects for both target and non-target attacks. Each attack uses no less than 1000 adversarial samples to calculate the attack success rate. The white-box adversarial sample generation schemes, FGSM [4], BIM [7], and PGD [8], are implemented using the advertorch [3].

The overall of ICTA is described in Algorithm 1.

Algorithm 1 Training algorithm of our ICTA.

Inputs: Target model T; Training data $\mathcal{X}$; Number of classes C;
Output: Substitute model S;

Initialization: Set the parameters: $\lambda_1 = \lambda_3 = 1, \lambda_2 = 10$;
for x, y in $\mathcal{X}$ **do**
  for m =1 to M **do**
    $\phi \sim N(0, 3)$ // Sample random disturbance strength
    $y_{arb} \in \{1, 2, ..., C | y_{arb} \neq y\}$ // Choose class different from the ground truth;
    $x^m_{pes} = x - \phi \nabla_x L(S(x), y_{arb})$
    $L^m_{RP} = \mathcal{D}_{rp}(T(x^m_{pes}), S(x^m_{pes}))$
  end for
  $L_{KD} = \mathcal{D}_{kd}(T(x), S(x))$
  $L_{all} = \lambda_1 L_{KD} + \lambda_2 \sum_{m=1}^{M} L^m_{RP} + \lambda_3 L_{CE}$,
  Optimizing substitute model S based on loss L_{all};
end for
return S

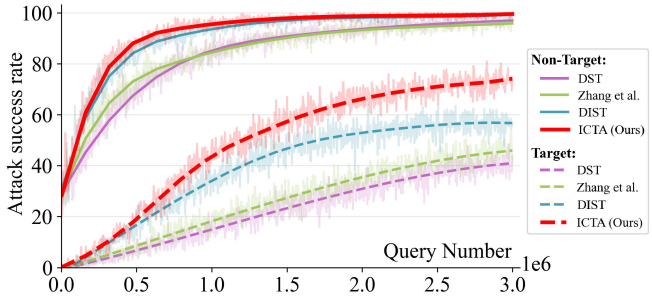

Fig. 7: Attack success rate comparison in data-free scenario.

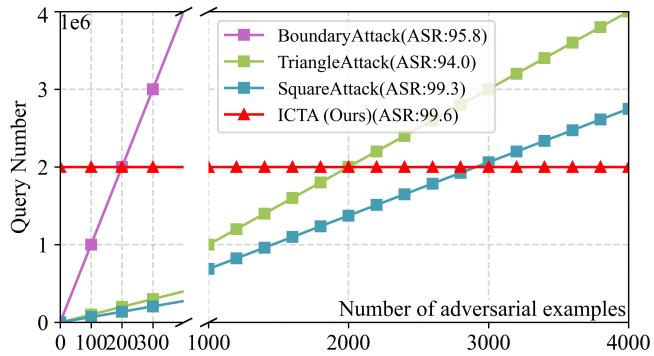

Fig. 8: Query number comparison with query-based attacks. The comparison schemes include BoundaryAttack [2], TriangleAttack [10], and SquareAttack [1].

B Fairer Comparisons in Data-free Scenario.

we conduct fairer comparison using the diffusion model as generator and MSE as our distance metric. The experimental setup is conducted under a data-free scenario, utilizing ResNet50 as the target model and selecting PGD as the white-box adversarial example generation method. As in Fig. 7, our ICTA exhibits superior attack success rates in both target and non-target attacks under the data-free scenario, with particularly significant improvements in target attacks. These experimental results indicate that our approach can substantially enhance the similarity between the substitute model and the target model.

C Query comparison with query-based attacks.

We compare the query number with query-based attacks at similar attack success rates in Fig. 8. It is observed that the query number of query-based attacks increases proportionally with the number of adversarial examples. In contrast, our ICTA requires a fixed query number (2M) and generates adversarial examples

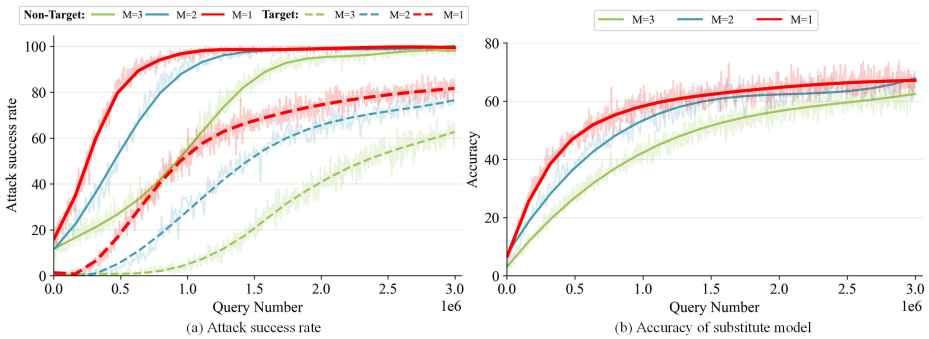

Fig. 9: The ablation experiment on the number of positional exploration samples (*i.e.* M in Eq. (9)). PGD [8] is used as the default white-box adversarial sample generation scheme and ResNet50 [5] on CIFAR100 [6] is used as the target model.

directly without additional queries. If a larger number of adversarial examples (more than 3000) are generated, our ICTA requires fewer queries than query-based attacks, and the advantage grows as the number increases.

D Additional analysis of Sec. 4.3

In Table 5, in some cases, the RP is greater than 1.0. We attribute this to the relatively low attack accuracy of FGSM. As a simple and fast technique, FGSM adds only one perturbation to generate adversarial examples. This leads to higher randomness and fluctuations of adversarial examples, e.g., originally inaccurate adversarial examples may successfully attack target model. In contrast, PGD and BIM use multiple iterations to ensure the precise perturbation and thus have stable and higher attack accuracy.

E Ablation on the Number of PES

We conduct ablation experiment on the number of Position Exploration Sample (PES), *i.e.* M in Eq. (9), as shown in Fig. 9. It can be observed that the efficiency is optimal when $M = 1$, *i.e.* only one PES is generated for each clean sample. The primary reason is that, in the early stages of substitute training, the substitute model is not well established, resulting in limited effectiveness of the generated PES. Consequently, the increase in the number of PES will only lead to more queries to the target model, thereby reducing efficiency.

F Ablation on Parameters of the PES

The ablation experiment on the parameters settings of the PES are given in Tab. 8, which contains several possible parameters. 'Fixed 1' represents a fixed

Table 8: The ablation comparison of the parameters in PES. 'Fixed 1' indicates that the noise strength is fixed to 1, 'Uniform 1' and 'Uniform 3' denote that the noise strength obeys a uniform distribution $U(0,1)$ and $U(0,3)$, 'Normal 1' and 'Normal 3' indicate that the noise strength obeys a Gaussian distribution $N(0,1)$ and $N(0,3)$. The best result are masked as **bolded**, and the sub-optimal results are masked as underlined.

	Target			Non-target		
Method	FGSM	BIM	PGD	FGSM	BIM	PGD
Fixed 1	8.12	58.08	64.38	**86.82**	98.80	99.28
Uniform 1	6.35	66.96	74.81	85.38	98.14	98.59
Uniform 3	7.06	69.11	76.58	85.45	98.86	99.38
Normal 1	6.61	68.62	75.54	84.84	98.59	98.91
Normal 3	**8.52**	**72.79**	**79.34**	86.76	**99.22**	**99.50**

noise strength of 1, 'Uniform 1' denotes a noise strength set to a uniform distribution $U(0,1)$ with a minimum value of 0 and a maximum value of 1, 'Uniform 3' indicates a noise strength set to a uniform distribution $U(0,3)$, 'Normal 1' represents a noise strength set to a Gaussian distribution $N(0,1)$ with a mean of 0 and a variance of 1, and 'Normal 3' indicates a noise strength set to a Gaussian distribution $N(0,3)$. ResNet50 [5] on CIFAR100 [6] is used as the target model.

It can be seen that a fixed noise strength is not sufficient to comprehensively explore the decision space. Additionally, the performance of uniformly distributed noise strength is not as expected, since a lot of PES is used to explore distant positions in the decision space. In contrast, when Gaussian distributed noise strength is used, PES is more distributed near the original class for meaningful exploration, and a small amount of PES can also be used to explore distant positions in the decision space. Our approach performs optimally or sub-optimally when the noise strength is set to a Gaussian distribution $N(0,3)$, and thus it is used as the preferred configuration for our method.

G Limitation

Although ICTA has made some progress in improving the effectiveness of black-box adversarial attacks, it faces the limitation in terms of computational complexity. This is limited by the nature of the substitute attacks, which require high computational complexity due to the need to optimize the model parameters, especially on large datasets and complex models. This may limit the feasibility of ICTA in practical applications.

Negative Impacts

This study emphasizes that attacks on neural networks are high probability to be successful even in a black-box environment. This type of research could be illegally used to attack the deployed neural networks, thereby generating adversarial samples that may influence the decision of the network.

Ethical Statements

This research adheres to ethical principles and is committed to applying the results in a legal, ethical and responsible manner to improve system security. In addition, we will actively take steps to ensure the safety and sustainability of our research and to maintain the trust of the scientific community and society.

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 (Jul 2020) 2
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 (Feb 2018) 2
3. Ding, G.W., Wang, L., Jin, X.: AdverTorch v0.1: An adversarial robustness toolbox based on pytorch. arXiv preprint arXiv:1902.07623 (2019) 1
4. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014) 1
5. 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) 3, 4
6. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Tront (2009) 3, 4
7. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016) 1
8. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2018) 1, 3
9. Meng, L., Shao, M., Wang, F., Qiao, Y., Xu, Z.: Advancing Few-Shot Black-Box Attack With Alternating Training. IEEE Transactions on Reliability pp. 1–15 (2024) 1
10. 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. Springer (2022) 2