# Exploring Effective Data for Surrogate Training Towards Black-box Attack (Appendix)

Xuxiang Sun Gong Cheng Hongda Li Lei Pei Junwei Han
School of Automation, Northwestern Polytechnical University, Xi'an, China
{xuxiangsun,hongda,peilei}@mail.nwpu.edu.cn {gcheng,jhan}@nwpu.edu.cn

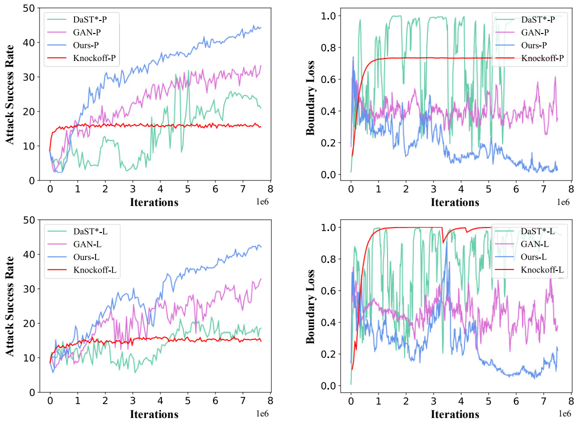



Figure 1. **Iterations vs. Attack Success Rate (Untargeted) and Boundary Loss (BL) on CIFAR-100 dataset [3].** Here, BL denotes the distance from the synthesized data to the decision boundary. "*-P" denotes the probability-only scenario and "*-L" represents the label-only scenario.

## A. Extended Description for Our Method

In this section, we offer our readers the extended description for the proposed loss $\mathcal{L}_{\text{div}}$, i.e., when the batch-size $B$ of the input noise less than $C-1$, where $C$ is the class number of the dataset on which the victim model is deployed.

When it comes to $B < C-1$, it is obvious that we can not explore all the class regions by those samples. In this scenario, a less-than-ideal alternative is that we can make those samples as diverse as possible. That is to say, we can focus on pushing them towards the decision boundary of their own class and the other $B$ classes. Hence, the optimization function $\mathcal{L}_{\text{div}}$ then becomes Eq. (1), as shown in the following:

$$\mathcal{L}_{\text{div}} = \frac{1}{C}\sum_{k}^{C}\text{STD}[\sum_{j}^{B} Norm(g(\text{S}(x_j^k))_{\phi_k})]. \qquad (1)$$

Here, $g(*)_{\phi_k}$ represents calculating the softmax outputs of $*$, where the entries with their index fall into the set of $\phi_k$ are excepted. The expression for $\phi_k$ is shown below:

$$\phi_k = Sort_{C-B-1}\{Norm(\sum_{j}^{B}\text{S}(x_j^k)_k)\} \cup \{k\}, \qquad (2)$$

where $Sort_{C-B-1}\{*\}$ denotes taking the first $C-B-1$ entries of $*$ in ascending order and $\text{S}(*)_k$ represents taking all the entries of $\text{S}(*)$, of which the $k$-th entry is excepted.

## B. Evaluation Metrics

In this section, the calculation of the two evaluation metrics will be further introduced. Briefly speaking, we follow the protocol of the state-of-the-art [8].

First, let us denote the adversarial perturbations in the corresponding adversarial examples are $\epsilon$. The adversarial examples with $||\epsilon|| < 8$ are seen as the ***valid*** ones. Then, the Attack Success Rate (ASR) is calculated by $n/m$ . Here, $n$ is the number of ***valid*** adversarial examples that can fool the victim model.

For $\text{ASR}_{\text{untar}}$, $m$ is the number of the images that are classified correctly by the victim model, and $n$ is the number of the ***valid*** adversarial examples that can be classified to any other class except its original one by the victim model. For $\text{ASR}_{\text{tar}}$, $m$ is the number of the images that are not classified to the specific target labels, and $n$ is the number of the ***valid*** adversarial examples that can be classified to the specific target labels by the victim model.

## C. Extended Results

### C.1. Extended Curves on CIFAR-100

**Implementation Details.** For the experiments in this part, all the compared methods are trained for 75 epochs with the default learning rates on each dataset. The learning rates for all the methods are fixed during training. Besides, for the experiments on CIFAR-10 dataset [3], we use VGG-16 [7] as the surrogate model and ResNet-18 [2] as the victim model. While for CIFAR-100 dataset [3], both the surrogate model and the victim model are VGG-19 [7]. Moreover, the

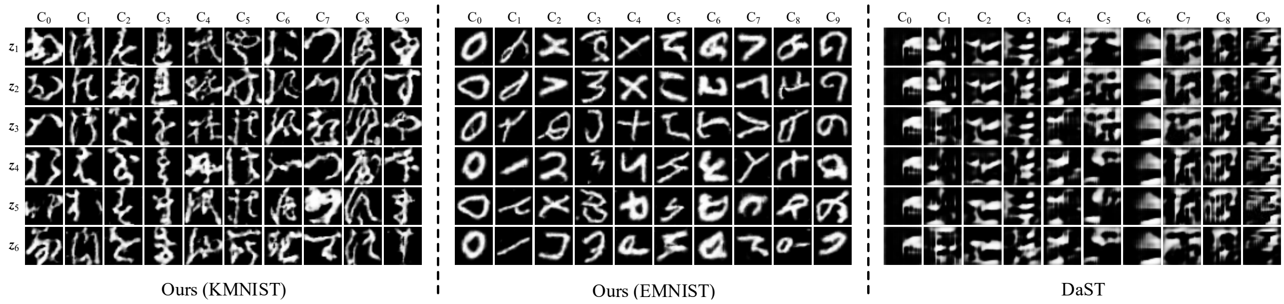



Figure 3. **Visualization of the synthesized data generated by our method (the specific proxy dataset is shown in bracket) and DaST [11] on MNIST dataset [4].** Here, $z_i$ represents different input noise, and $C_i$ donotes different class.

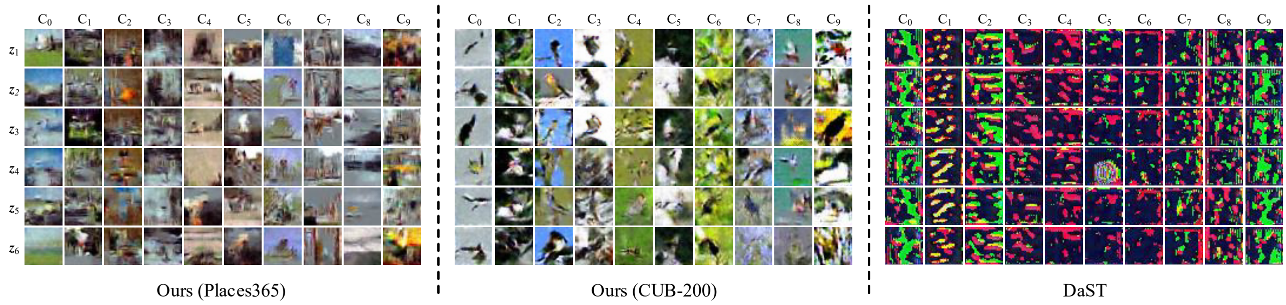
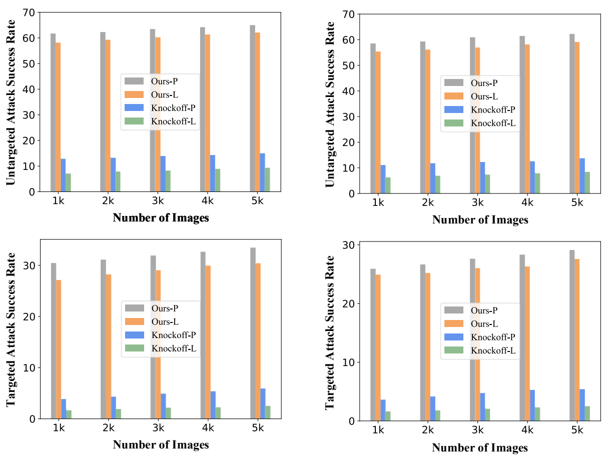



Figure 4. **Visualization of the synthesized data generated by our method (the specific proxy dataset is shown in bracket) and DaST [11] on CIFAR-10 dataset [3].** Here, $z_i$ represents different input noise, and $C_i$ donotes different class.




(a) Places365 dataset

(b) CUB-200 dataset

Figure 2. **Data ablation studies on CIFAR-100 dataset with (a) Places365 [10] and (b) CUB-200 [9].** Here, we set the number of proxy images between 1k and 5k to evaluate the performances of our method and Knockoff under two attack scenarios. The victim model is ResNet-50 [2].

method "GAN" in Fig. 1 represents using the vanilla adversarial loss [1] to train the proposed generator and discriminator in this paper. Associating Fig. 1 in this part with Fig. 1 in the main paper, we can find that for most cases, high boundary loss failed to provide sustainable improvement. Besides, it is astonishing that using GAN directly without any model-specific constraint can even exceed DaST [11], and the values of boundary loss in all the cases are lower than DaST [11]. Moreover, when the boundary loss of our approach get increased abnormally, the attack performance will decrease sharply.

Based on this observation, as we mentioned in the main paper, it is reasonable to make a conjecture that samples lay close to the decision boundary may be effective relatively. In addition, the distribution built by the proxy images indeed contains a large proportion of the samples that are effective for surrogate training. Even if there is no specific constraint, the generator can still find those samples. However, considering the inherently instability [1, 6] of GAN, we can not just rely on it to search the effective samples for surrogate training. To this end, our approach adds two losses to search the specified samples that are effective for efficient surrogate training.

## C.2. Data Ablation Studies on CIFAR-100

Here, we report the data ablation studies on CIFAR-100 [3] with the two proxy datasets. The victim model here is ResNet-50 [2]. From Fig. 2 we can find that our method is not much sensitive to the size of the proxy data. It is worth emphasizing that although the size of proxy images has no impressive impact on the performance of both our method and Knockoff [5], we think the reasons for the two cases are quite different. For our method, we argue this can be attributed to the distributions established by proxy images with the size between 1k to 5k are not much different. While for Knockoff [5], we believe that the reason is the underfitting problem still dominates, *i.e.*, the proxy images with a size lower than 5k are unavailing that can not cause significant improvement. As a result, our approach can make full use of the proxy images without the risk of underfitting due to insufficient data.

### C.3. Visualizations of the Synthesized Data

In the end, we provide the visualization of the data synthesized by DaST [11] and our method, respectively. For our method, the synthesized data via two proxy datasets are all exhibited. Here, Fig. 3 and Fig. 4 offer the visualizations on two typical victim datasets, *i.e.*, MNIST [4] and CIFAR-10 [3] dataset. Looking through Fig. 3 and Fig. 4, we can see that the intra-class diversity of the synthesized data via our method are generally larger than those via DaST. Besides, associating the Kernel Density Estimation curves of the main paper with Fig. 4, we can find that yet the inter-class similarity of the synthesized data via our method is large, images belong to different classes still be visually distinguishable and stylistically similar. That is, the synthesized samples are label-controllable with large inter-class similarity, as expected. While for DaST, the label-controllable property is relatively poor, *i.e.*, samples belonging to different classes are almost visually identical. Moreover, since we utilize a discriminator to limit the searching space of the generator to the distribution established by the proxy images, it is obvious that the synthesized samples own high semantic similarity with the proxy dataset. Thus, according to the results of peer comparison, we think that the semantic content of the synthesized data may be unimportant for the surrogate training. In turn, the class-specific properties (*e.g.*, the inter-class similarity and the intra-class diversity) of those synthesized samples may provide sound development.

## References


[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inform. Process. Syst.*, 2014. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Int. Conf. Comput. Vis.*, pages 770–778, 2016. 1, 2

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Tech Report*, 2009. 1, 2

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proc. IEEE*, pages 2278–2324, 1998. 2

[5] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4954–4963, 2019. 2

[6] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proc. Adv. Neural Inform. Process. Syst.*, volume 29, pages 2234–2242, 2016. 2

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Represent.*, 2015. 1

[8] Wenxuan Wang, Bangjie Yin, Taiping Yao, Li Zhang, Yanwei Fu, Shouhong Ding, Jilin Li, Feiyue Huang, and Xiangyang Xue. Delving into data: Effectively substitute training for black-box attack. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4761–4770, 2021. 1

[9] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 2

[10] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2017. 2

[11] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 234–243, 2020. 1, 2