Enhancing targeted transferability via suppressing high-confidence labels: supplementary material

Hui Zeng, Tong Zhang, Biwei Chen, and Anjie Peng

The supplementary document consists of three parts of content: A) Ablation study on N_h and T; B) Image quality comparison; C) Iterative vs. generative attacks.

A Ablation study on N_h and T

1) **Influence of the number of high-confidence labels**. We study the influence of N_h on the transfer success rate in the single-model, random-target transfer scenario, with the source model fixed as DenseNet121. As shown in Fig. 1(a), the optimal N_h values for different source-target model pairs are between 5 and 15. This can be explained as follows. A very small N_h cannot suppress enough high-confidence labels. On the other hand, the gradients associated with different labels may contradict each other when N_h is large. Hence, there is a trade-off in setting N_h. In our study, we set $N_h = \mathbf{10}$. It is also observed from Fig. 1(a) that the choice of $N_h = 10$ is strictly better than that of $N_h = 0$, which verifies the necessity of suppressing high-confidence labels. Note $N_h = 0$ means suppressing the original label y_o only (Eq. (7) of the paper).

2) **Influence of the timing T of introducing the orthogonal gradient**. Next, we study the influence of T on the transfer success rate in the attack scenario as above. In this experiment, we fix the other parameters of the proposed method and let T vary from 0 to 1. As shown in Fig. 1(b), the optimal T varies between different target models. The main takeaway is that the orthogonal gradient should not be introduced at the attack's very beginning or end. This phenomenon can be explained as follows. On the one hand, the high-confidence labels calculated at the beginning of the attack differ significantly from the high-confidence labels of the final AEs. On the other hand, $T = 1$ means the orthogonal gradient is not introduced. Based on the above considerations, we set $T = \mathbf{0.75}$ in this study.

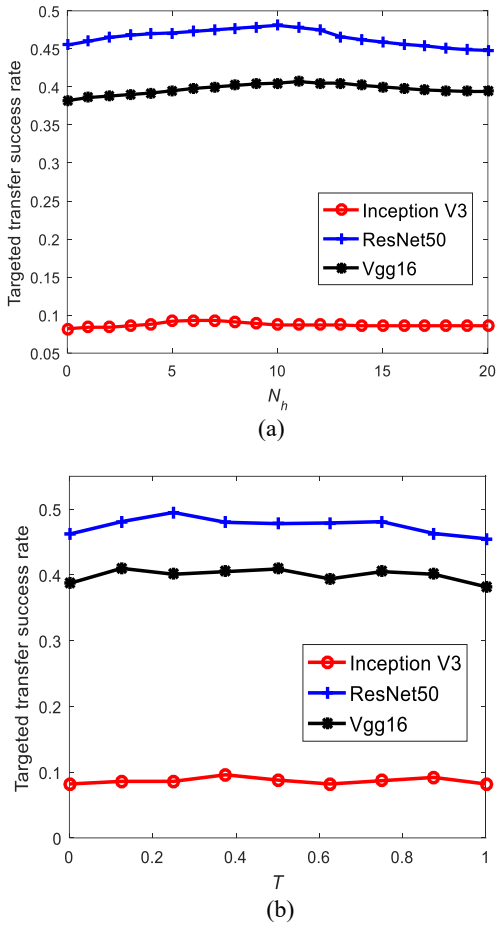


Fig. 1. Ablation study on N_h and T. The source model is a pretrained DenseNet121 model. (a) Targeted transfer success rate as a function of N_h. (b) Targeted transfer success rate as a function of T.

B Image quality comparison

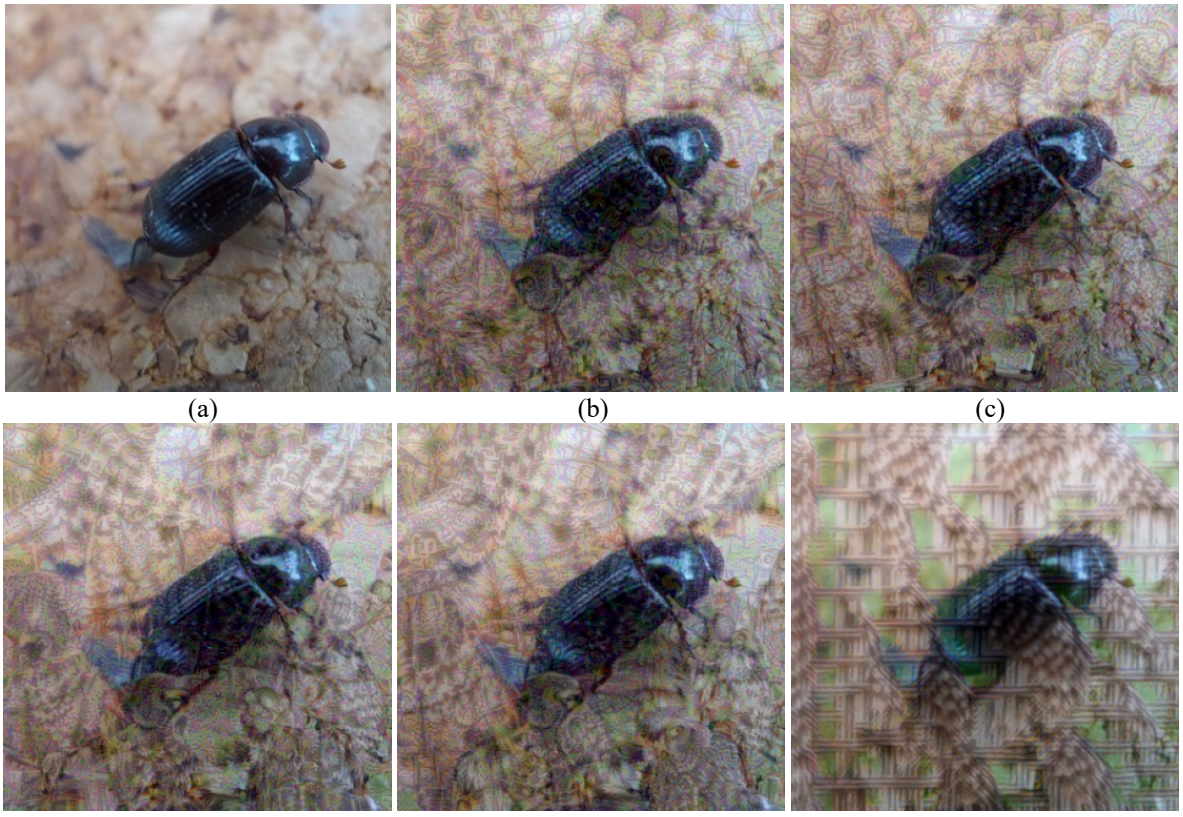
Besides attack ability, image quality is also crucial for a successful adversarial example in practice. Note the distortion introduced on AEs may vary significantly for different methods, even restricted to the same L_∞ norm.

We first show AEs generated with compared methods in Fig. 2. While the perturbation introduced by the iterative methods resembles noise, that introduced by TTP shows a weaving-like pattern. Hence, although TTP-generated AE has the highest PSNR value, it is the most suspicious under human inspection.

The average PSNR value of the AEs for different attacks is provided in Table 1. Under the same perturbation budget, all the iterative attacks introduce almost the same amount of noise to the image.

C Iterative vs. generative attacks

We compare the iterative attacks with a representative generative attack TTP for different perturbation budgets. Since



(a) (b) (c)

(d) (e) (f)

Fig. 2. The visual comparison of the AEs generated by different methods, $\epsilon = 16$. The source model is Res50, and the target class is 'great grey owl'. (a) Original image, (b) CE, *PSNR*=24.33, (c) Po+Trip, *PSNR*=24.31, (d) Logit, *PSNR*=24.44, (e) proposed, *PSNR*=24.45, (f) TTP, *PSNR*=25.42.

Table 1. Image quality comparison in terms of averaged PSNR in the single-model scenario. The source model is Res50.

perturbation budget	CE	Po+Trip	Logit	Proposed	TTP
$\epsilon = 8$	30.32	30.31	30.34	30.35	30.77
$\epsilon = 16$	24.44	24.42	24.51	24.51	25.46

training a dedicated GAN generator for each source model and each target class is prohibited for us, we download ten pretrained generators and follow the '10-Targets' setting of TTP. The source model is Res50, and Inc-v3, Dense121, and VGG16 acted as target models. Fig. 3 shows competitors' averaged attack success rates (over three black-box models and ten target classes). TTP is advantageous in large-budget scenarios, whereas iterative attacks retain better attack ability in low-budget cases. Such results may help us in choosing different schemes for different purposes.

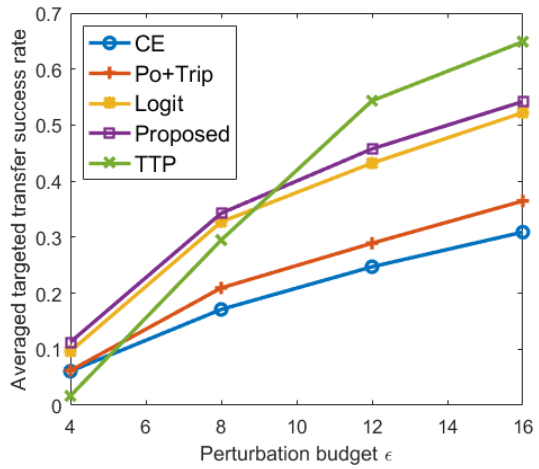


Fig. 3. Targeted transfer success rate of TTP vs. iterative attacks, averaged over 3 models and 10 target classes