Two Heads Are Better Than One: Averaging along Fine-Tuning to Improve Targeted Transferability: supplementary material

Hui Zeng, Sanshuai Cui, Biwei Chen, and Anjie Peng

The supplementary document consists of five parts of content: A) Ablation study on the decaying factor γ; B) Visualization of FFT and AaF in a 2D subspace; C) Attack performance on Swin; D) Attack performance in the most difficult-target scenario; E) Visual comparison.

A Ablation study on γ

We make an ablation study on the newly introduced decaying factor γ of AaF in the random-target scenario. The source models are Inc-V3, Res50, Dense121, and VGG16, the same as our paper. The targeted success rates are averaged over three hold-out models and a VIT-based model, Swin. We let γ vary from 0 to 1 with a step of 0.1. Note γ reduces to the vanilla FFT method with $N_{ft} = 15$, and $\gamma = 1$ means a simple average over the fine-tuning trajectory. Fig. 1 shows that the optimal γ for different source models vary. Generally, $\gamma \in [0.4, 0.8]$ will be a good choice for all surrogates. In our study, we experimentally set $\gamma = 0.8$ for simplicity.

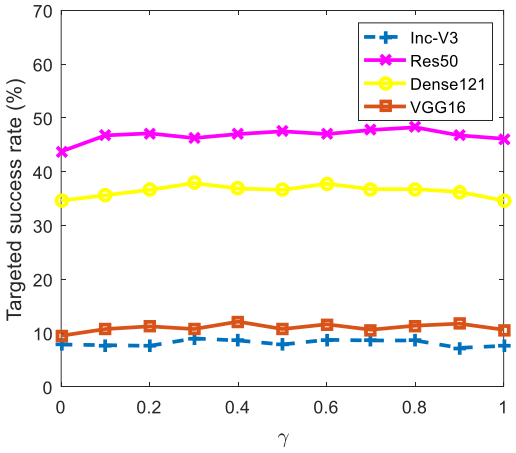


Fig. 1. Effect of γ on AEs' transferability averaged over three hold-out models and Swin. The baseline attack is Logit.

B Visualization of FFT and AaF in a subspace

Besides the examples in the Fig. 3 of the paper, we provide additional examples here. To avoid the bias introduced by cherry pick, we investigate the first 20 samples of the ImageNet-compatible dataset. The planes are generated using the following steps: First, we use the AEs $\boldsymbol{I}'$, $\boldsymbol{I}'_{FFT}$ and $\boldsymbol{I}'_{AaF}$ to span a 2D subspace. Then we calculate the logit of a point $\boldsymbol{I}'_{sample}$ in the spanned subspace based on an ensemble of models:

$$logit_{ensemble} = 1/N(\sum_{i}^{N} logit_{model,i}(\boldsymbol{I}'_{sample})),$$

where N is the model number and $logit_{model,i}()$ is the logit output w.r.t. the target class of the i-th model. The value of $logit_{ensemble}$ indicates the targeted transferability of $\boldsymbol{I}'_{sample}$. As shown in Fig. 2, in most cases, the proposed AaF method steers AE towards a more central region than FFT.

C Attack performance on Swin

Table 1 reports the targeted transferability of the VIT-based model Swin in both random-target and most difficult-target scenarios. The proposed AaF shows superiority to existing fine-tuning schemes in most cases. The only exception is when the IncV3 is the surrogate, whereas the plain FFT sometimes performs best.

D Attack to the most difficult target

Table 2 compares different fine-tuning schemes in the most difficult-target scenarios. Compared to the results reported in Table I of the paper, the improvement from fine-tuning is more remarkable under the most difficult target scenario. Let Res50 be the surrogate, Logit+AaF improves the Logit attack by 18.2% (55.5% vs. 47.0%) in the random-target scenario and 24.2% (38.8% vs. 31.3%) in the most difficult-target scenario, in terms of targeted transfer rate averaged over three victim models.

E Visual comparison

We visually compare AEs crafted with different methods in Fig. 3. The target label for all samples is 'hippopotamus.' The perturbations introduced by TTP are more suspicious under human inspection, while those introduced by the iterative methods resemble noise.

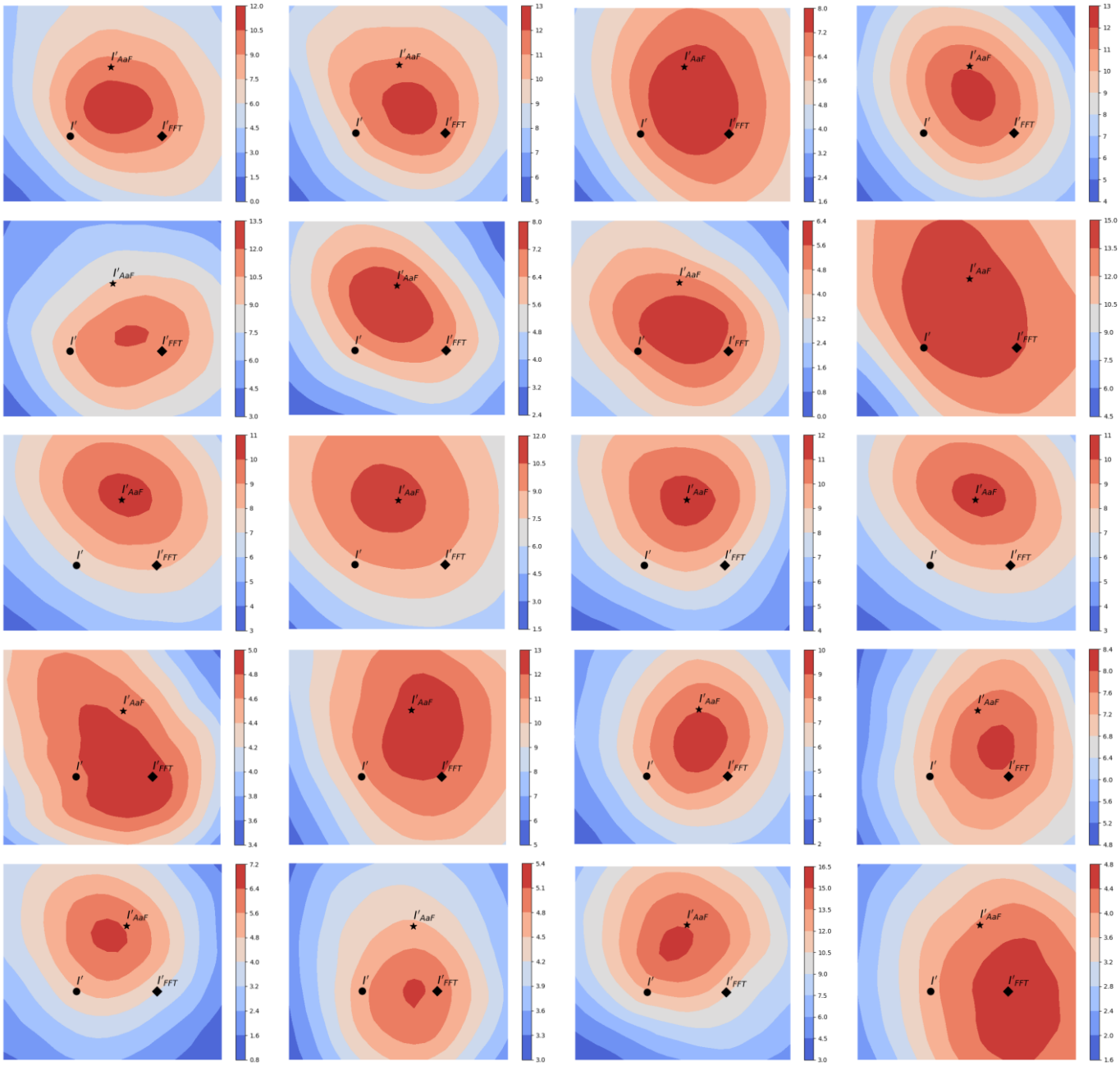


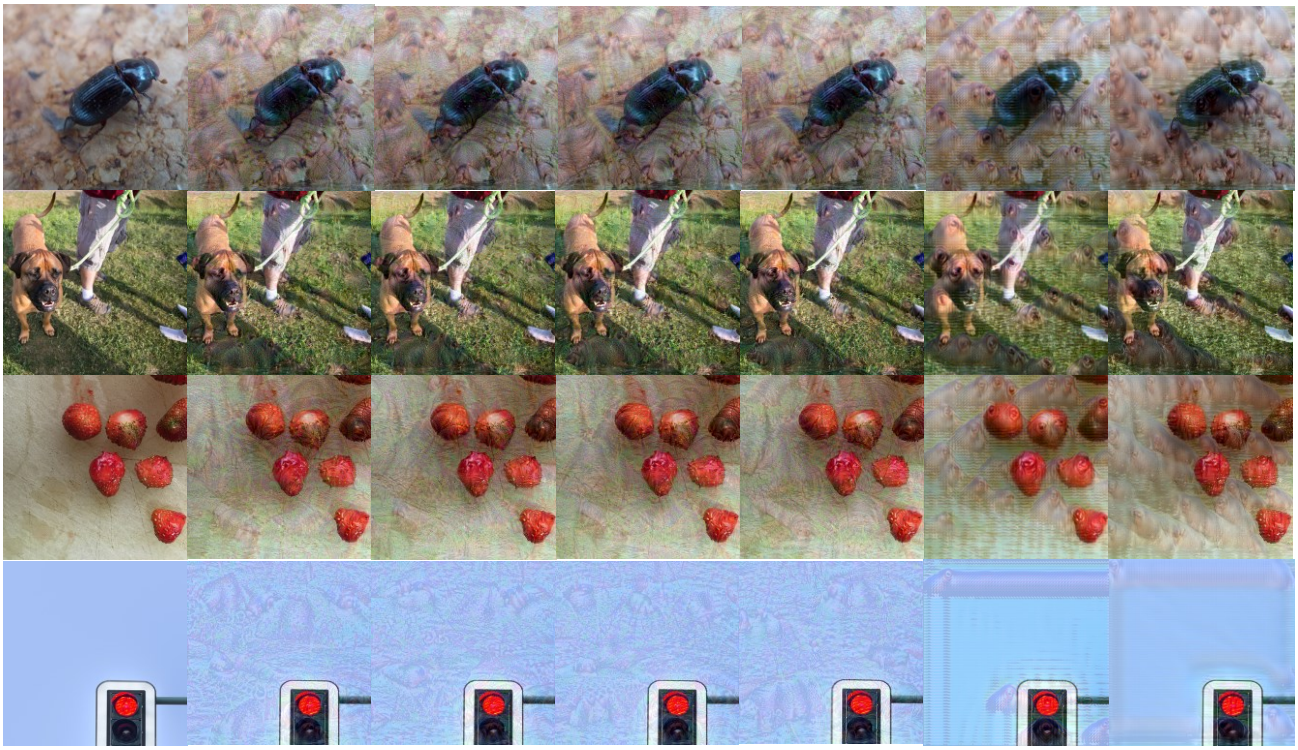
Fig. 2. The logit plane of AEs with different fine-tuning schemes.The baseline attack is Logit. The AEs are crafted against a Res50 model, and the logits are calculated based on an ensemble of models: IncV3, Dense121, and VGG16.

Table 1. Targeted transfer success rate (%): no fine-tuning/ fine-tuning with ILA/FFT/AaF. The target model is Swin.

	Attack	Source Model: Res50	Dense121	VGG16	Inc-v3
random-target	CE	5.1/7.5/7.7/**11.5**	1.7/3.6/4.3/**9.2**	0.0/0.4/0.5/**0.7**	0.0/0.1/0.3/**0.7**
	Logit	13.4/20.3/18.9/**22.1**	10.5/12.4/12.6/**13.4**	6.2/7.6/8.6/**9.5**	0.2/1.6/**1.8**/1.6
	Margin	16.5/17.3/21.7/**24.1**	11.5/13.6/14.5/**16.1**	6.4/7.4/7.8/**8.3**	0.9/**1.4**/1.1/1.1
	SH	17.1/**24.7**/22.4/23.6	9.3/14.4/15.7/**17.6**	7.1/8.7/9.2/**10.1**	0.2/1.7/**1.9**/1.8
	SU	21.3/22.9/20.8/**24.2**	12.9/15.1/15.4/**16.2**	8.1/8.3/**9.9**/9.5	0.9/1.1/1.2/**1.5**
most difficult-target	CE	3.0/3.5/4.2/**5.7**	1.2/2.8/3.1/**6.1**	0.0/0.0/0.0/0.0	0.4/0.8/1.0/**1.8**
	Logit	9.2/8.9/12.1/**13.4**	4.7/7.5/7.9/**9.1**	5.0/5.5/5.6/**6.1**	0.5/0.9/1.2/**1.6**
	Margin	10.2/10.4/11.9/**13.3**	5.0/7.5/7.7/**8.5**	6.5/6.1/6.2/**6.5**	0.2/**0.4**/0.3/0.3
	SH	8.1/10.2/11.4/**11.7**	5.2/7.5/8.6/**10.7**	3.1/3.3/3.5/**3.8**	0.3/0.5/1.1/**1.3**
	SU	13.2/13.5/13.6/**13.9**	8.9/10.2/10.7/**12.0**	4.5/4.8/4.9/**6.2**	0.9/1.0/1.3/**1.4**

Table 2. Targeted transfer success rate (%): no fine-tuning/ fine-tuning with ILA/FFT/AaF, in the most difficult-target scenario.

Attack	Source Model: Res50 →Inc-v3	→Dense121	→VGG16	Source Model: Dense121 →Inc-v3	→Res50	→VGG16
CE	1.3/2.5/3.1/**9.8**	25.8/44.8/45.3/**53.9**	15.0/30.6/29.7/**41.6**	1.2/4.5/6.1/**9.0**	6.5/19.6/23.4/**34.8**	3.6/14.7/19.2/**29.6**
Logit	3.6/7.3/7.5/**10.2**	51.6/56.6/53.1/**59.7**	38.6/45.3/44.3/**46.6**	3.5/7.6/8.3/**10.2**	22.7/38.8/41.6/**45.1**	18.3/31.5/37.5/**42.2**
Margin	4.0/7.5/8.4/**8.6**	52.5/60.7/57.1/**62.5**	38.0/51.8/47.6/**53.1**	4.0/9.5/8.7/**10.3**	24.5/43.7/44.4/**49.2**	18.0/36.9/35.0/**40.6**
SH	4.0/7.0/8.1/**9.8**	54.5/60.6/57.9/**63.3**	41.6/49.7/51.2/**56.2**	4.0/8.1/8.6/**10.5**	24.5/41.9/43.3/**43.8**	21.2/37.2/40.4/**41.1**
SU	5.0/6.8/7.5/**12.3**	56.2/55.9/54.8/**61.4**	44.1/48.3/49.7/**53.6**	4.4/7.6/8.0/**9.6**	27.4/41.2/41.7/**46.2**	24.3/37.7/37.6/**40.1**
Attack	Source Model: VGG16 →Inc-v3	→Res50	→Dense121	Source Model: Inc-v3 →Res50	→Dense121	→VGG16
CE	0.0/0.0/0.0/0.0	0.0/0.5/1.4/**1.8**	0.0/0.4/0.6/**1.9**	2.4/3.9/5.9/**7.8**	4.2/4.9/7.8/**10.3**	2.3/4.6/5.0/**8.1**
Logit	0.3/0.4/0.4/**0.7**	5.6/8.0/8.9/**12.6**	7.0/8.8/8.5/**12.2**	3.8/6.1/8.1/**10.4**	5.5/7.2/10.5/**11.8**	3.2/6.1/7.9/**8.4**
Margin	0.0/0.3/0.4/**0.5**	4.4/6.4/6.6/**9.2**	6.0/8.6/**10.5**/**10.5**	2.5/5.2/7.5/**9.2**	3.5/7.1/8.8/**12.4**	2.0/4.3/6.2/**9.1**
SH	0.1/0.2/0.3/**0.5**	3.9/7.6/8.1/**9.5**	6.8/8.6/9.4/**9.7**	3.0/5.5/8.2/**10.1**	4.9/7.4/10.6/**12.8**	3.4/4.9/7.2/**10.0**
SU	0.1/0.1/0.2/**0.8**	5.7/7.1/7.6/**9.2**	7.4/8.2/9.7/**10.4**	3.9/5.8/8.1/**10.5**	6.7/8.9/12.1/**14.3**	3.9/5.3/9.6/**11.2**



(a) (b) (c) (d) (e) (f) (g)

Fig. 3. The visual comparison of the AEs generated by different methods, $\epsilon = 16$. The target class is '*hippopotamus*'. (a) Original image, (b) Logit+*AaF*, (c) Margin+*AaF*, (d) SupHigh+*AaF*, (e) SU+*AaF*, (f) TTP, (g) C-GSP.