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 six parts of content: A) The pseudo-code of the proposed method; B) Ablation study on the decaying factor γ; C) Visualization of FFT and AaF in a 2D subspace; D) Attack performance against transformer-based models; E) Attack performance in the most difficult-target scenario; F) Visual comparison.

A Pseudo code

Due to page limitation, we provide the pseudo-code of the proposed AaF attack in **Algorithm 1**.

B 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 $\gamma = 0$ 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.

Algorithm 1 AaF attack

```
Input: A benign image $\boldsymbol{I}$ with original label $y_o$; target label $y_t$.
Parameter: Iterations $N$ of the baseline attack; Fine-tuning iterations $N_{ft}$; decaying factor $\gamma$.
Output: Adversarial image $\boldsymbol{I'}_{aaf}$.
Mount a baseline attack for $N$ iterations, and obtain an AE $\boldsymbol{I}'$.
Calculate aggregate gradient $\bar{\Delta}_k^{I',t}$ from $\boldsymbol{I}'$, and $\bar{\Delta}_k^{I,o}$ from $\boldsymbol{I}$.
Obtain the combined aggregate gradient $\bar{\Delta}_k^{combine}$ as (1).
$\boldsymbol{I}'_{aaf,0} = \boldsymbol{I}'$
for $t$ = 1 to $N_{ft}$ do
    Fine-tune $\boldsymbol{I}'_{aaf,t-1}$ with the optimization objective defined in (4) and obtain $\boldsymbol{I'}_{ft,t}$.
    $\boldsymbol{I}'_{aaf,t} = \gamma \boldsymbol{I}'_{aaf,t-1} + \boldsymbol{I'}_{ft,t}$
end for
return $\boldsymbol{I}'_{aaf,N_{ft}}$.
```

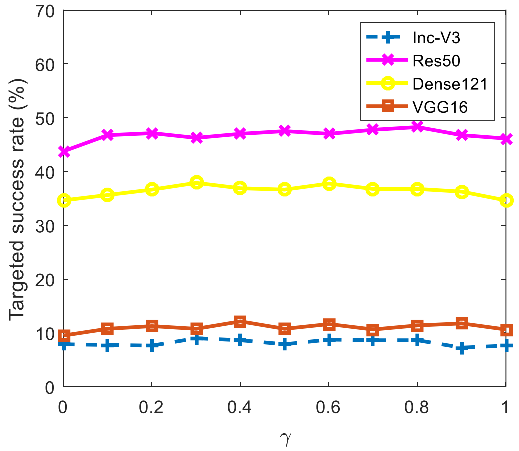

Fig. 1. Effect of γ on AEs' transferability averaged over three hold-out models and Swin. Each curve represents a different surrogate. The baseline attack is Logit.

C 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:

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

where N is the model number and $logit_i()$ 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.

D Attack performance on transformers

Table 1 reports the targeted transferability against four transformer-based models, Swin, vit_b_16, pit_b_24, and visformer, in the random-target scenario. Compared to the

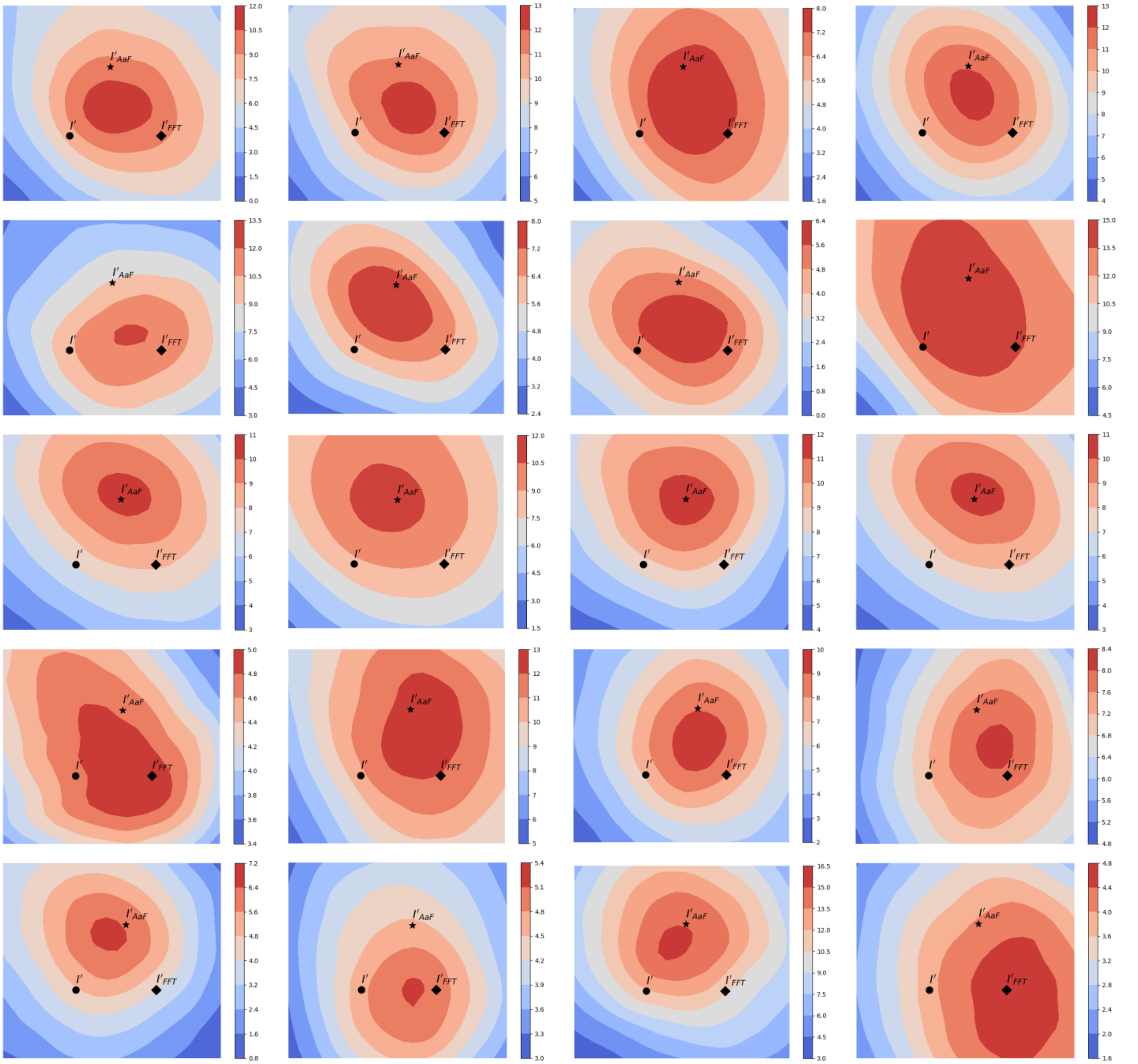

Fig. 2. The logit plane of AEs with different fine-tuning schemes.The baseline attack is Logit. The AEs are crafted against a Res50 model, and the logits are calculated based on an ensemble of models: IncV3, Dense121, and VGG16.

results reported in Table I of the paper, transferring from CNNs to transformers is much more challenging than transferring between CNNs. Nevertheless, in most cases, the proposed AaF is superior to existing fine-tuning schemes. For example, when Res50 is the surrogate, the average success rate of CE+AaF surpasses CE+FFT (the second best) by 46.8% (6.9% vs. 4.7%).

E Attack to the most difficult target

Table 2 compares different fine-tuning schemes in the most difficult-target scenarios. Compared to Table I of the paper, the improvement from fine-tuning is more remarkable under the most challenging 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.

F 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.

Table 1. Targeted transfer success rate (%) against transformers in the random-target scenario. No fine-tuning/ fine-tuning (ILA/FFT/AaF).

Source Model	Attack	Victim Model Swin	 vit_b_16	 pit_b_24	 visformer	AVG
Res50	CE	5.1/7.5/7.7/**11.5**	0.6/0.8/1.8/**2.7**	2.0/2.2/2.2/**3.2**	4.8/8.0/7.0/**10.1**	3.1/4.6/4.7/**6.9**
	Logit	13.4/20.3/18.9/**22.1**	2.7/3.8/5.3/**5.5**	6.0/6.7/6.3/**8.0**	16.0/20.2/**20.3**/19.7	9.5/12.8/12.7/**13.8**
	Margin	16.5/17.3/21.7/**24.1**	4.8/5.8/6.0/**6.4**	7.6/**9.9**/9.0/9.7	19.5/22.5/23.2/**24.4**	12.1/13.9/15.0/**16.2**
	SH	17.1/**24.7**/22.4/23.6	3.7/3.4/5.8/**6.2**	7.3/**9.2**/8.7/8.8	20.1/22.2/22.4/**23.9**	12.1/14.9/14.8/**15.6**
	SU	21.3/22.9/20.8/**24.2**	5.0/4.1/5.0/**5.2**	4.8/5.0/**6.8**/6.5	20.0/18.9/18.9/**20.6**	12.8/12.7/12.9/**14.1**
Dense121	CE	1.7/3.6/4.3/**9.2**	1.2/1.6/2.2/**3.4**	1.2/2.2/2.3/**4.0**	6.2/11.4/13.2/**16.9**	2.6/4.7/5.5/**8.4**
	Logit	10.5/12.4/12.6/**13.4**	2.5/5.1/4.6/**5.2**	4.7/6.5/6.6/**7.4**	23.5/**29.4**/28.9/29.1	10.3/13.4/13.2/**13.8**
	Margin	11.5/13.6/14.5/**16.1**	3.6/5.3/6.4/**6.5**	5.2/7.0/7.1/**8.0**	20.8/28.0/28.8/**31.7**	10.3/13.5/14.2/**15.6**
	SH	9.3/14.4/15.7/**17.6**	2.9/5.0/**5.2**/4.9	4.0/**7.2**/6.2/7.0	25.2/29.8/30.3/**31.9**	10.4/14.1/14.4/**15.4**
	SU	12.9/15.1/15.4/**16.2**	4.4/4.4/4.5/**5.3**	4.4/4.9/4.9/**5.1**	23.9/**27.0**/25.2/26.3	11.4/12.9/12.5/**13.2**
VGG16	CE	0.0/0.4/0.5/**0.7**	0.0/0.0/0.0/**0.2**	0.0/0.0/0.1/**0.4**	0.6/0.6/0.7/**2.2**	0.2/0.3/0.3/**0.9**
	Logit	6.2/7.6/8.6/**9.5**	0.2/0.5/**0.8**/0.6	1.4/1.9/**3.0**/2.3	6.7/7.7/**7.9**/6.8	3.6/4.4/**5.1**/4.8
	Margin	6.4/7.4/7.8/**8.3**	0.1/0.4/0.3/**0.5**	1.8/1.8/2.6/**2.7**	4.2/4.2/**6.2**/5.5	3.1/3.5/4.2/**4.3**
	SH	7.1/8.7/9.2/**10.1**	0.4/0.4/**1.0**/0.9	2.0/1.7/**2.9**/2.5	9.4/8.7/11.3/**11.5**	4.7/4.9/6.1/**6.3**
	SU	8.1/8.3/**9.9**/9.5	0.8/**0.9**/0.8/0.7	2.2/2.8/2.6/**2.9**	12.7/11.8/11.4/**13.1**	6.0/6.0/6.2/**6.6**
Inc-v3	CE	0.0/0.1/0.3/**0.7**	0.2/0.2/0.2/**0.8**	0.2/0.4/0.2/**0.8**	0.7/0.8/0.6/**1.8**	0.3/0.4/0.3/**1.0**
	Logit	0.2/1.6/**1.8**/1.6	0.3/0.5/**1.2**/1.1	0.6/0.7/0.7/**0.9**	1.0/1.0/**1.2**/1.1	0.4/1.0/**1.2**/**1.2**
	Margin	0.9/**1.4**/1.1/1.1	0.4/0.4/0.5/**0.7**	0.4/0.5/0.6/**0.8**	0.9/1.6/1.2/**1.6**	0.7/1.0/0.9/**1.1**
	SH	0.2/1.7/**1.9**/1.8	0.2/0.5/**0.8**/0.7	0.7/0.9/**1.0**/0.9	0.8/0.2/1.0/**1.2**	0.5/0.8/**1.2**/**1.2**
	SU	0.9/1.1/1.2/**1.5**	0.2/0.4/**0.6**/0.5	0.2/0.3/0.4/**0.6**	1.0/0.9/0.9/**1.1**	0.6/0.7/0.8/**0.9**

Table 2. Targeted transfer success rate (%): no fine-tuning/ fine-tuning (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	4.1/7.5/8.4/**8.6**	52.7/60.7/57.1/**62.5**	38.0/51.8/47.6/**53.1**	4.0/9.5/8.7/**10.3**	24.7/43.7/44.4/**49.2**	18.2/36.9/35.0/**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**

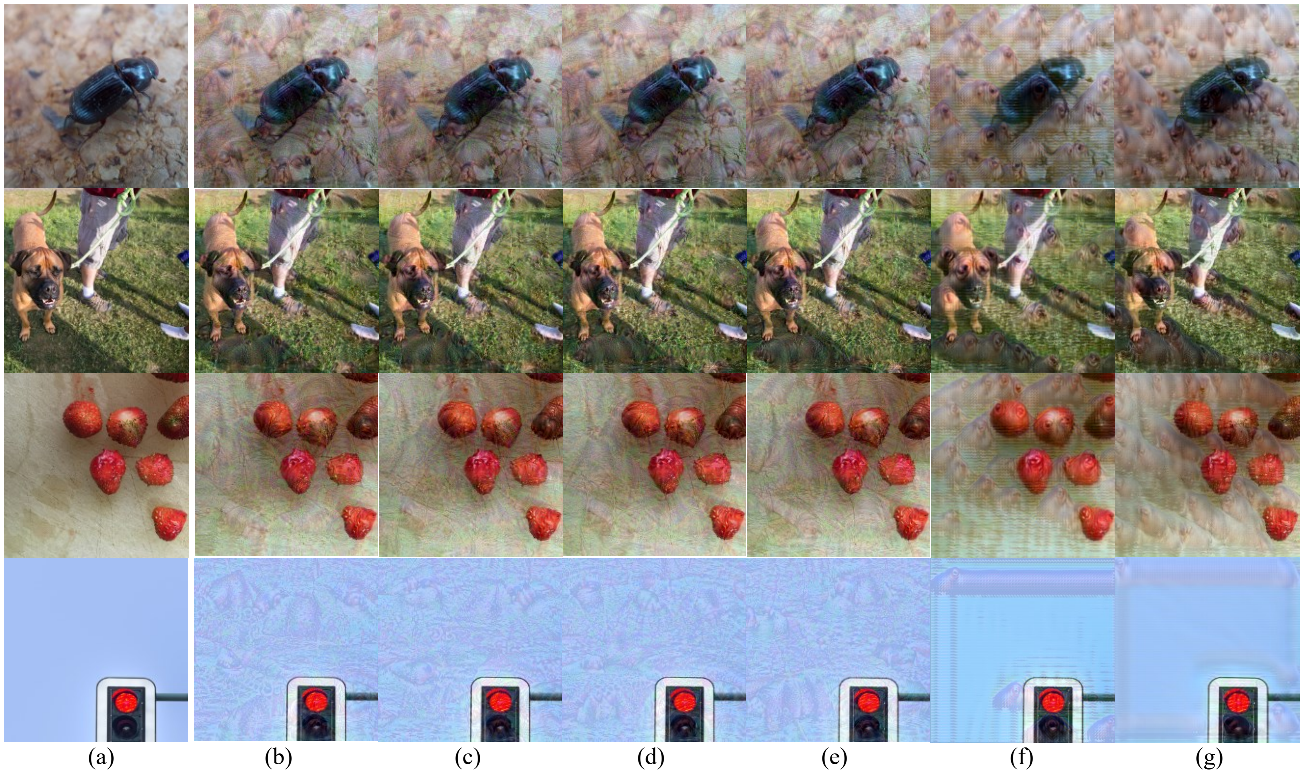

Fig. 3. The visual comparison of the AEs generated by different methods, $\epsilon = 16$. The target class is '*hippopotamus*'. (a) Original image, (b) Logit+*AaF*, (c) Margin+*AaF*, (d) SupHigh+*AaF*, (e) SU+*AaF*, (f) TTP, (g) C-GSP.