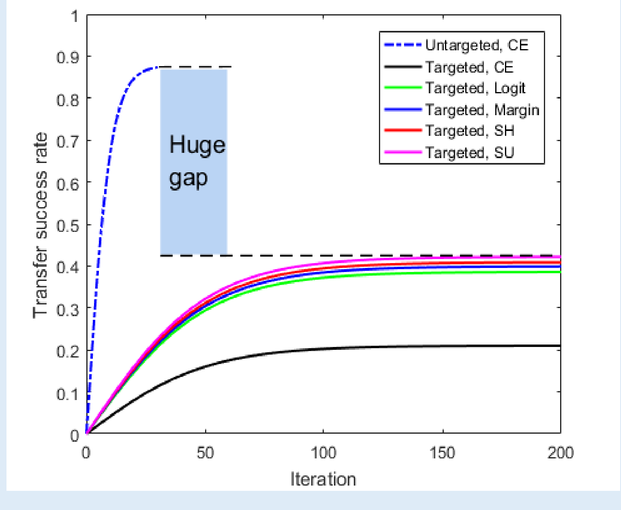
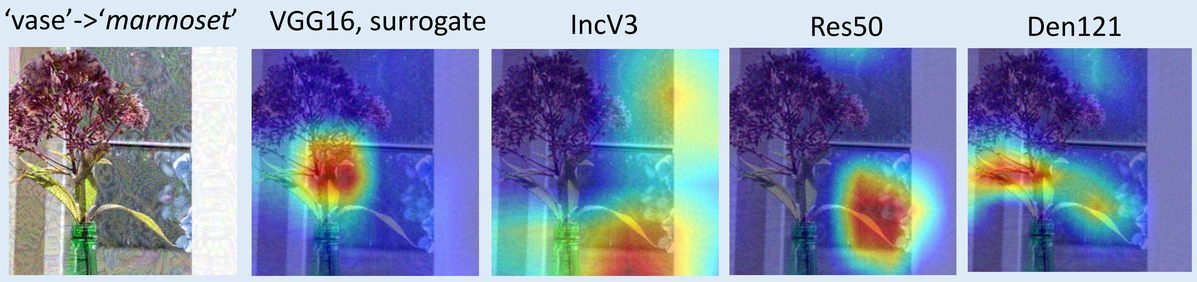
Everywhere attack: Attacking locally and globally to boost targeted transferability

Hui Zeng, Sanshuai Cui, Biwei Chen and Anjie Peng

BACKGROUND

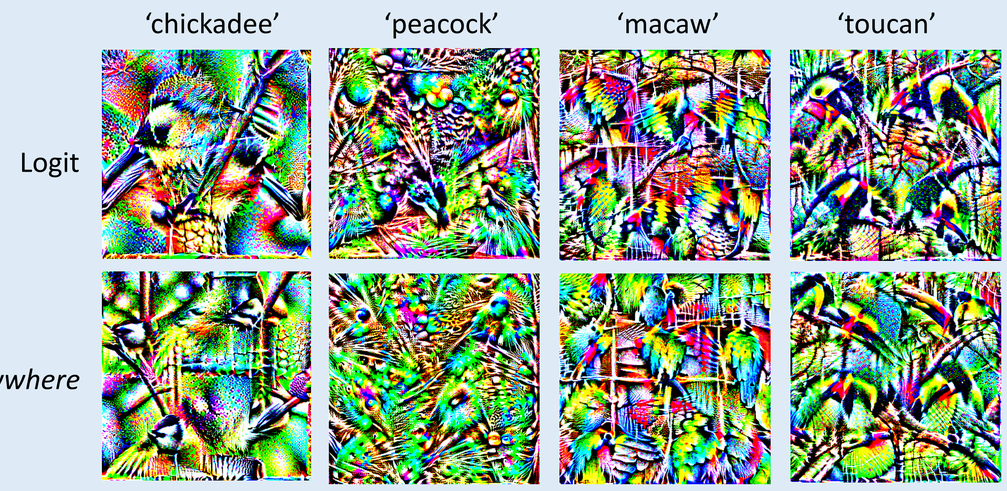
Targeted transferability is much more challenging than its untargeted counterpart. We believe that the crux lies in inconsistent attention regions when different DNN models identify a counterfactual, targeted class from a given image. In the example below, a '*vase*' image is attacked to '*marmoset.*' The surrogate model focuses on the lower area of the flower crown (the second image from left) in synthesizing a '*marmoset.*' In contrast, victim models pay attention to strikingly different regions in recognizing it.



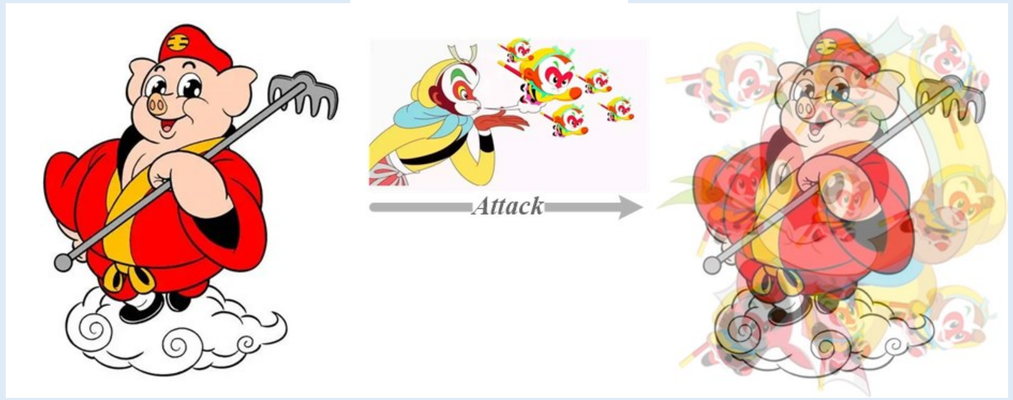


QUALITATIVE COMPARISON

Data-free Targeted UAPs crafted with Res50adv. Compared to the baseline attack, the *everywhere* attack tends to plant more target objects with smaller sizes into the obtained UAP.

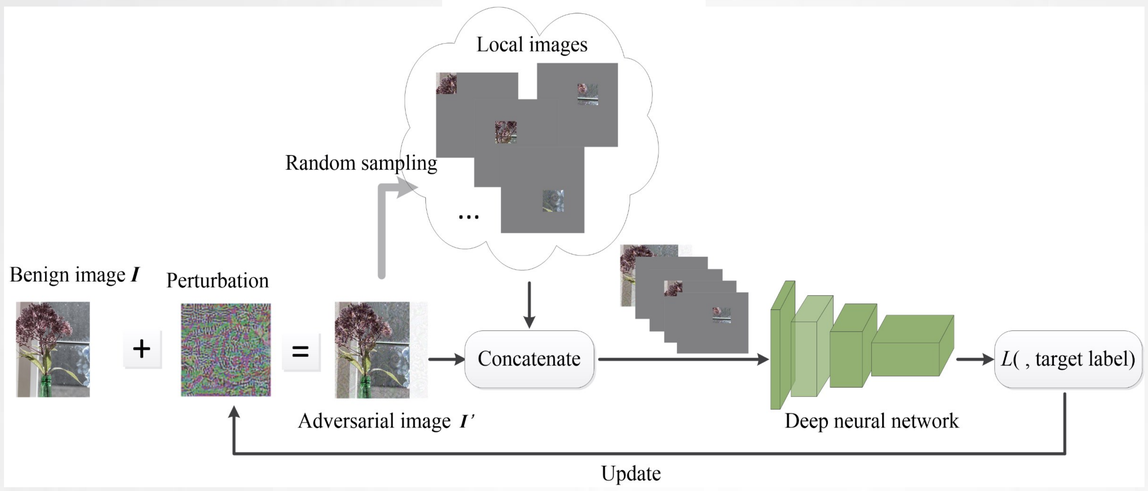


MOTIVATION



In the Chinese novel "*Journey to the West*", the main character Wukong (the monkey) often gains an advantage in fights through cloning spells, as opponents are confused about which one is his true self. Inspired by this, we hypothesis that increasing the number of target objects is helpful for transferability. To this end, we split a victim image into non-overlap blocks and jointly mount a targeted attack on each block.

ALGORITHM



RESULTS

Dataset: 1000 images of 299 × 299 pixels from the ImageNet-compatible dataset.
Networks: Inceptionv3, Resnet50, DenseNet 121, VGG16bn, Swin_t.
Competitors: CE, Logit [Zhao, 2021NIPS], Margin [Weng, 2023TIFS], SH [Zeng, 2023ICIP], SU [Wei, 2023CVPR] and CFM [Byun, 2023CVPR].
The table below provides targeted transfer success rate (%) w.o. /w. the proposed *everywhere* scheme.
Try it yourself: https://github.com/zengh5/Everywhere_Attack

	Source model: Res50					Source model: Dense121				
Attack	→IncV3	→Den121	→VGG16	→Swin	AVG	→IncV3	→Res50	→VGG16	→Swin	AVG
CE	3.9/14.1	44.9/62.3	30.5/52.2	5.2/19.0	21.1/36.8	2.8/10.3	19.0/41.7	11.3/50.6	1.8/19.2	8.7/30.5
Logit	9.1/22.3	70.0/78.5	61.9/69.3	13.4/28.8	38.6/49.7	7.4/17.6	42.6/58.5	36.3/54.2	10.5/23.8	24.2/38.5
Margin	10.9/21.7	70.8/80.8	61.2/69.4	16.5/33.1	39.9/51.3	7.6/19.8	44.7/58.9	33.4/56.4	11.7/24.6	24.4/39.9
SH	9.9/17.8	74.2/82.7	62.5/78.2	17.1/37.3	40.9/54.0	8.7/12.9	47.4/64.3	40.5/64.1	9.3/23.6	26.6/41.2
SU	11.1/21.9	72.5/79.2	63.9/67.4	21.3/34.2	42.2/50.7	10.0/17.2	49.2/63.4	42.3/55.5	13.5/23.1	28.8/39.8
CFM	41.4/55.3	83.3/87.7	77.2/81.9	41.5/54.2	60.9/69.8	35.2/43.6	77.3/84.8	66.6/73.9	27.1/43.4	51.6/61.4