







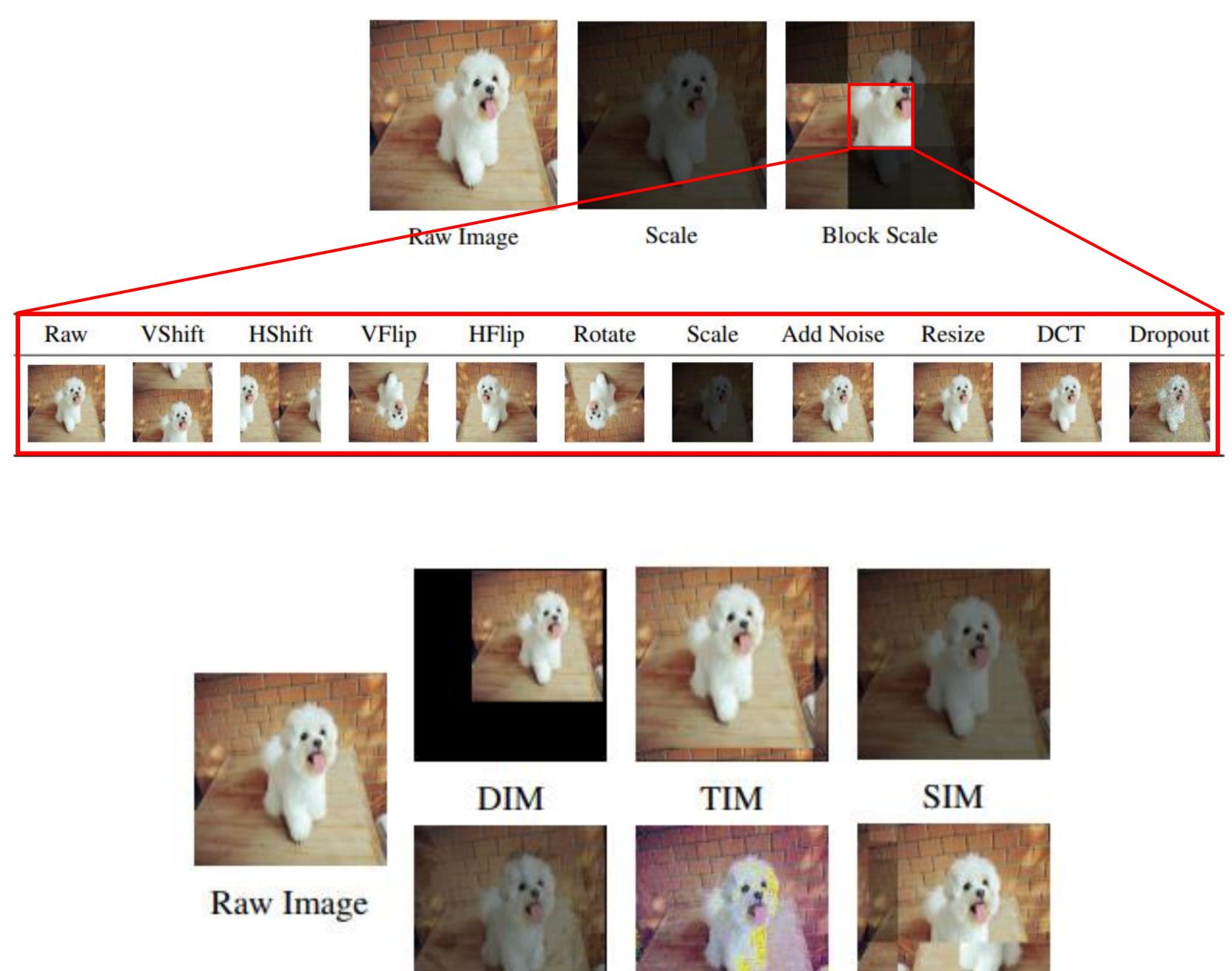
Motivation

Assumption: Without harming the semantic information, the more diverse the transformed image is, the better transferability the adversarial examples have.

	TIM	DIM	SIM	SSA	Admix		
Transferability LPIPS		77.6 0.43				$LPIPS(\boldsymbol{x}, \hat{\boldsymbol{x}}) =$	E

The Struture of Image

Defination: Given an image x, which is randomly split into s x s blocks, the relative relation between each anchor point is the structure of image, where the anchor point is the center of the image block.



Structure Invariant Transformation for better Adversarial Transferability ICCV23

Xiaosen Wang¹, Zeliang Zhang², Jianping Zhang³ ¹Huawei Singularity Security Lab & ²HUST & ³Chinese University of Hong Kong Paper, code, and data are available:

$$\overline{W} \sum_l \sum_{h,w} \|oldsymbol{z}_{h,w}^l - \hat{oldsymbol{z}}_{h,w}^l\|_2$$

SIA (Ours)

SSA

Admix

Structure Invaria

Algorithm 1: Structure Invariant At

Input: Classifier $f(\cdot)$ with the loss benign sample x with ground The maximum perturbation iterations T and decay facto number s; Number of trans

Output: An adversarial example. $1 \alpha = \epsilon/T$ $a_0 = 0$ $x_0^{adv} = x$

1
$$\alpha = \epsilon/T$$
, $g_0 = 0$, $x_0^{\alpha \alpha \beta} =$
2 for $t = 0 \rightarrow T - 1$ do

Constructing a set \mathcal{X} of N tran using SIT

$$\bar{g}_{t+1} = \frac{1}{N} \sum_{x_i \in \mathcal{X}} \nabla_a$$

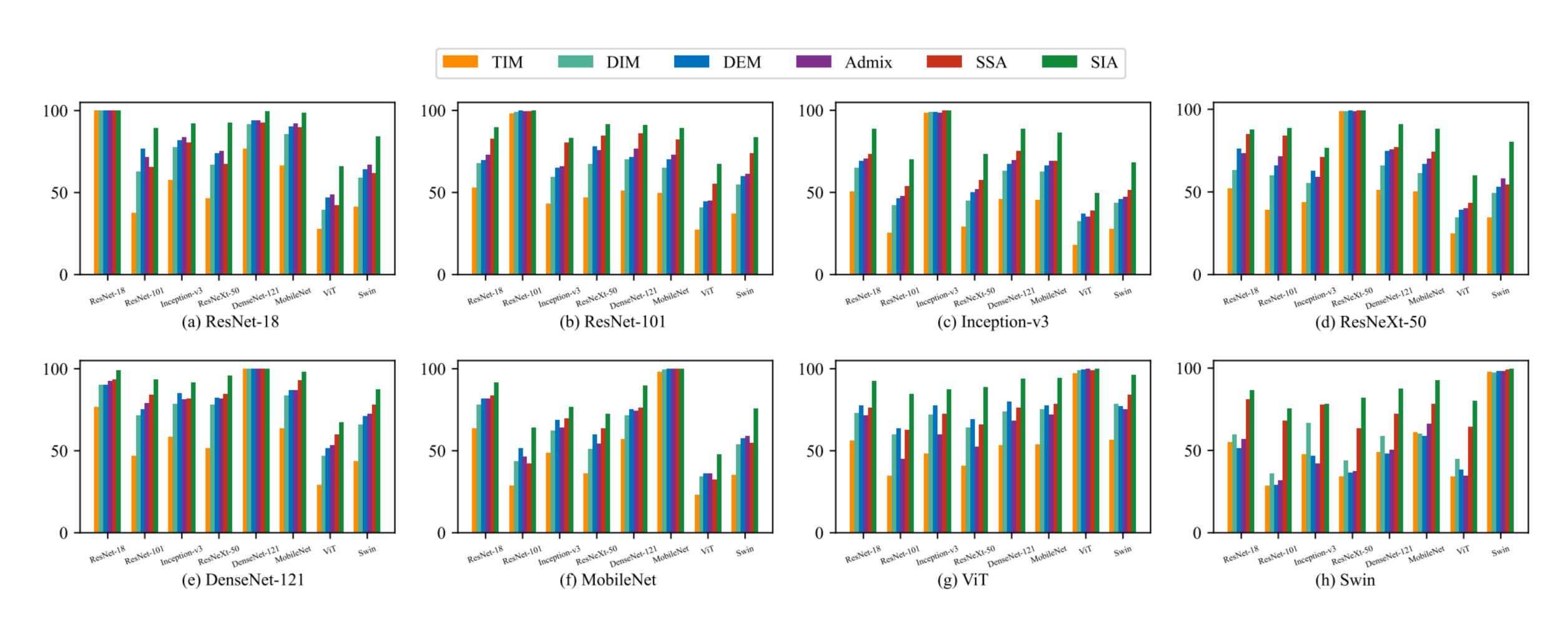
Updating the momentum:

$$\boldsymbol{g}_{t+1} = \mu \boldsymbol{g}_t + \frac{\boldsymbol{g}_t}{\|\boldsymbol{\bar{g}}\|}$$

Updating the adversarial examp

$$\boldsymbol{x}_{t+1}^{adv} = \operatorname{Clip}(\boldsymbol{x}_t^{adv} + \alpha \cdot \operatorname{sign})$$

7 return x_T^{adv}



1													
ttack	(T)	TIM	DIM	DEM	Admix	SSA S	SIA	TIM	DIM DE	M Admi	x SSA	SIA	
s function J and-truth lab a ϵ , number of or μ ; Splittin formed images	el y; of ng ges N	100 - 80 - 80 - 60 - 40 - 20 - 0 -					Attack success rates (%)	40 - 20 -					
	C	ResNet-18	ResNet-101 Inception-v3	ResNeXt-50 DenseNet-1	21 _{MobileNet} vi	r _{Swin}		IncRes-v2ens	HGD R&P NIPS-13	FD JPEG	Bit-Red RS	NRP	
nt on \mathcal{X} :			Ablation study										
$J(\boldsymbol{x}_i, y)$	(2)		F	ResNet-18 ——— Incept ResNet-101 ——— ResNet	tion-v3 ->- DenseNo eXt-50 - *- MobileN	et-121 + ViT let - Swin			ResNet-18 ResNet-101	Inception-v3 E ResNeXt-50 N	PenseNet-121		
~ (~1,3)	N -7			•				100 -			• • •		
		e attack success rates (%) - 0.6 - 0						success rates (%) - 06 - 08 - 07 - 08					
		- 08 CCess 12						- 08 Cess 13					
\bar{l}_{t+1}	(2)	ack suc						rck suc			_+ + +		
$t+1 \ _1$	(3)	ige atts						ge attack					
		- 09 Average						Average	7				
ole:		50 -	1	1	1			40 -					
(-)	1) (1)	1 2 3 4 5 12 4 6 8 10 12 14 16 18 20 22 24 26 28 30 Number of blocks Number of blocks											
$(g_{t+1}), 0, 1$	(4)	SIA	-VShift	-HShift	-VFlip	-HFlip	-Rotate	-Sclae	-Add Noise	-Resize	-DCT	-Dropo	
		92.1	89.7	90.1	90.1	88.4	88.3	90.1	90.6	90.2	90.1	90.7	

Attacking a Single Model



https://github.com/xiaosen-wang/SIT

