Improving Black-Box Generative Attacks via Generator Semantic Consistency

Jongoh Jeong[1], Hunmin Yang[1,2], Jaeseok Jeong[1], and Kuk-Jin Yoon[1]
[1]Visual Intelligence Lab., KAIST, [2]Agency for Defense Development

Project Page

Introduction

- Generative transfer attacks are efficient but remain **underexplored in how their internal representations** affect black-box transferability.
- We improve transfer by **enforcing semantic consistency** via EMA-based feature alignment, reducing semantic drift without adding inference cost.
- Our method integrates into existing generative attacks and **consistently boosts black-box transfer** across models, tasks, and domains.

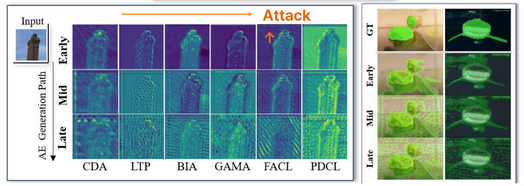
Attack	Input data aug.	Generator feature-level	Perturbed image-level	Surrogate mid-level layer feature-level	Surrogate output logit-level	GT label required?
GAP Poursaeed et al. (2018)	-	-	-	-	✔	✔/✗
CDA Wang et al. (2022)	-	-	-	-	✔	✔/✗
LTP Nakka & Salzmann (2021)	-	-	-	✔	-	-
BIA Zhang et al. (2022b)	-	-	✔	✔	-	-
GAMA Aich et al. (2022)	-	-	-	✔	-	✔
FACL Yang et al. (2024a)	✔	-	-	✔	-	-
PDCL Yang et al. (2024b)	-	-	-	✔	-	✔
Our focus	-	✔	-	-	-	-

Motivation

Research Questions

1. At **what stage** of synthesis do semantic cues **deteriorate**?
2. Which generator block(s) **most influence transferability**?

Closer look into perturbation generator



(a) Intermediate perturbations on the object ↑

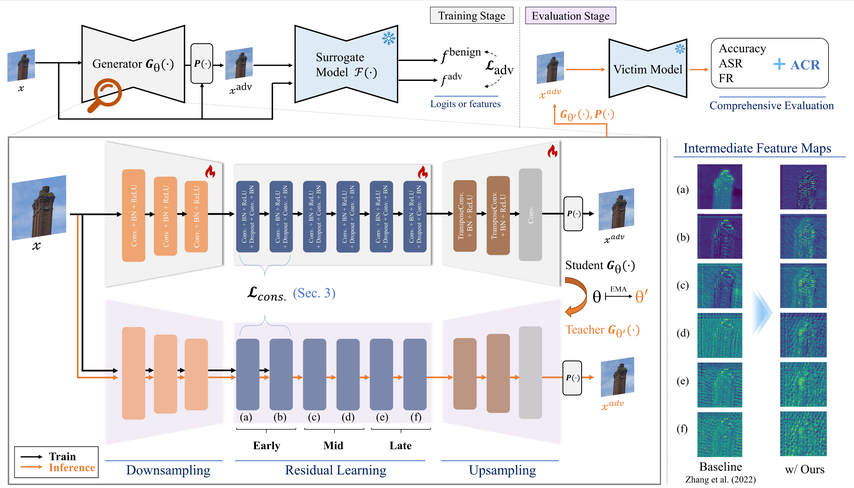
(b) More semantics reside in the early stage

Feature Cluster Foreground IoU	Intermediate block			Std.Dev. (Variability) ↓
Method	Early	Mid	Late	Baseline → w/ Ours
CDA	37.72	36.74	32.14	2.77 → **1.51**
LTP	32.48	28.16	28.16	2.59 → 2.98
BIA	36.17	33.79	30.20	2.82 → **2.06**
GAMA	36.55	35.95	31.57	2.46 → **1.41**
FACL	36.48	34.38	31.76	2.19 → **1.17**
PDCL	35.31	33.59	31.00	2.08 → **0.71**

(c) Semantic variability ↓

(a) Generator intermediate feature maps for each block partition
(b) Predicted foreground masks from intermediate feature clusters on ImageNet-S from the baseline (BIA)
(c) Quantified variability in foreground IoU

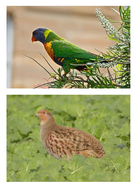
Semantically Consistent Generative Attack (SCGA)



Key Contributions

① Generator–internal evidence for perturbation semantics
② Generator–level semantic consistency guidance
③ Comprehensive evaluation with an added reliability measure

Novel metric for attack reliability: Accidental Correction Rate (ACR)



GT Label | Benign Prediction | Adv. prediction

Lorikeet — Ruffed Grouse — Lorikeet — **ACR**

Ruffed Grouse — Ruffed Grouse — Lorikeet — Conventional **ASR, FR**

Real-world examples:

Scenario #	GT Label	Benign pred.	Adv. pred.	Impact	Captured by
1	cat	cat ✔	cat ✔	Correct → Correct	Acc. only
2	cat	cat ✔	dog ✗	Correct → Incorrect	ASR, FR
3	van	truck ✗	bus ✗	Incorrect → Other incorrect	FR only
4	pelagic cormorant	albatross ✗	pelagic cormorant ✔	Incorrect → Correct	ACR, FR. Acc.

Reliable attack example:

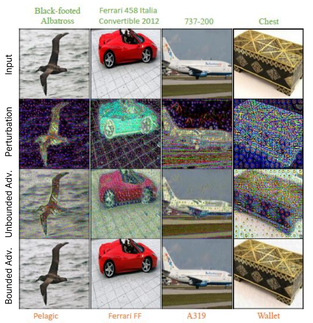
Cross-Setting	GT Label	Benign pred.	Intended Attack	Unreliable Attack
ImageNet → FGVC Aircraft	F-22 Raptor	F-22 Raptor ✗	F-18 Hornet ✗	F-22 Raptor ✔

Cross-	Domain	Model
Acc. ↓	47.10	44.13
ASR ↑	49.02	44.02
FR ↑	51.66	50.66
ACR ↓	9.66	8.32

Experimental Results

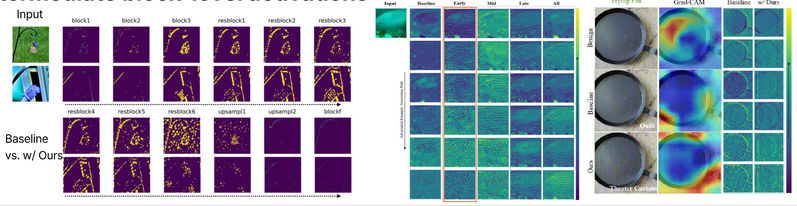
Main Results

Method	Cross-setting (Avg.)			
	Domain (Acc.)	Model (Acc.)	Task (SS; mIoU)	Task (OD; mAP50)
CDA	69.94	50.27	22.90	29.54
w/ Ours	**54.82**	**43.38**	**22.71**	**28.83**
LTP	49.91	48.33	25.34	25.90
w/ Ours	**40.76**	**41.99**	**24.48**	**24.52**
BIA	51.07	45.17	24.75	24.72
w/ Ours	**47.10**	**44.13**	**23.40**	**24.52**
GAMA	48.56	44.58	25.82	24.36
w/ Ours	**46.09**	**43.46**	**24.81**	**24.20**
FACL	44.05	42.00	25.08	24.43
w/ Ours	**41.78**	**40.92**	**24.20**	**23.97**
PDCL	43.91	42.84	25.24	24.93
w/ Ours	**43.06**	**42.77**	**24.20**	**24.20**



3 Domains **21 Models** (11 CNN, 6 ViT, 2 Mixers, 2 Vision Mamba)

Intermediate block-level activations



Robust to defenses?

Method	Metric	Defense	Input Proc.	Purification
Baseline	Acc. (%) ↓	54.03	37.08	72.89
	ASR (%) ↑	23.75	53.95	7.79
	FR (%) ↑	32.78	59.56	14.41
	ACR (%) ↓	9.51	8.36	11.02
w/ Ours	Acc. (%) ↓	51.80	34.05	72.63
	ASR (%) ↑	26.52	57.75	8.18
	FR (%) ↑	35.28	63.01	14.97
	ACR (%) ↓	9.01	7.80	11.21

Applicable to targeted attacks?

Model Target	24	99	245	344	471	555	661	701	802	919	Avg.
DenseNet121											
M3D Zhao et al. (2023a)	71.24	77.12	73.35	85.37	71.75	73.79	60.63	78.70	69.66	17.34	67.90
w/ Ours	76.53	77.11	82.64	87.74	82.81	78.72	74.12	78.29	79.43	47.20	**76.46**
ResNet50											
M3D Zhao et al. (2023a)	68.54	75.48	77.67	79.43	77.64	80.05	54.48	89.04	55.85	9.51	66.77
w/ Ours	71.02	73.60	80.69	88.04	87.64	83.88	68.97	84.88	69.60	45.36	**75.37**

Method	Victim model						Avg.
	VGG16	GoogLeNet	Inc-v3	Res152	Dense121	Inc-v4	IncRes-v2	
CGNC Fang et al. (2024a)	14.71	2.03	2.77	2.68	8.31	2.41	0.96	4.84
w/ Ours	47.50	7.90	10.96	12.24	31.29	13.36	3.98	**18.18**