Improving Black-Box Generative Attacks via Generator Semantic Consistency

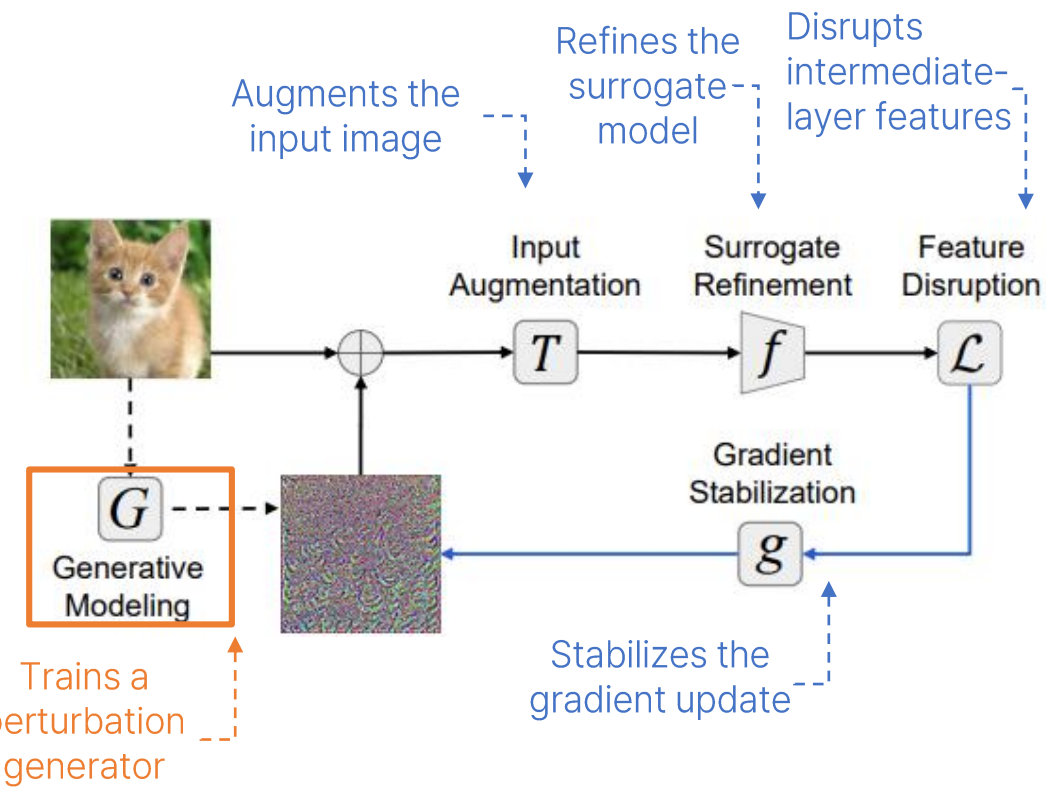
Jongoh Jeong[1], Hunmin Yang[1,2], Jaeseok Jeong[1], and Kuk-Jin Yoon[1,†]

[1]Visual Intelligence Lab., KAIST [2]Agency for Defense Development [†]Corresponding author

Background: Transferable Adversarial Attacks

- **Transfer-based Attacks** target black-box models by crafting adversarial examples on a local surrogate and transferring them to an unknown target
  - **Aim**: to make perturbations
    - less surrogate-specific, and
    - more generalizable across model, domain, defense settings
  - **Practical Advantages**
    - Requires no query interaction with the target
    - Cost-efficient alternative to conventional white-box iterative attacks such as PGD
  - **Core Objectives**
    - Transferability
    - Stealthiness (Imperceptibility)

Types of Transfer-based Attacks



Zhao, Zhengyu, et al. "Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights." IEEE T-PAMI 2025.

Fig. source: Revisiting Transferable Adversarial Images (TPAMI 2025)

Background: Generative Attacks

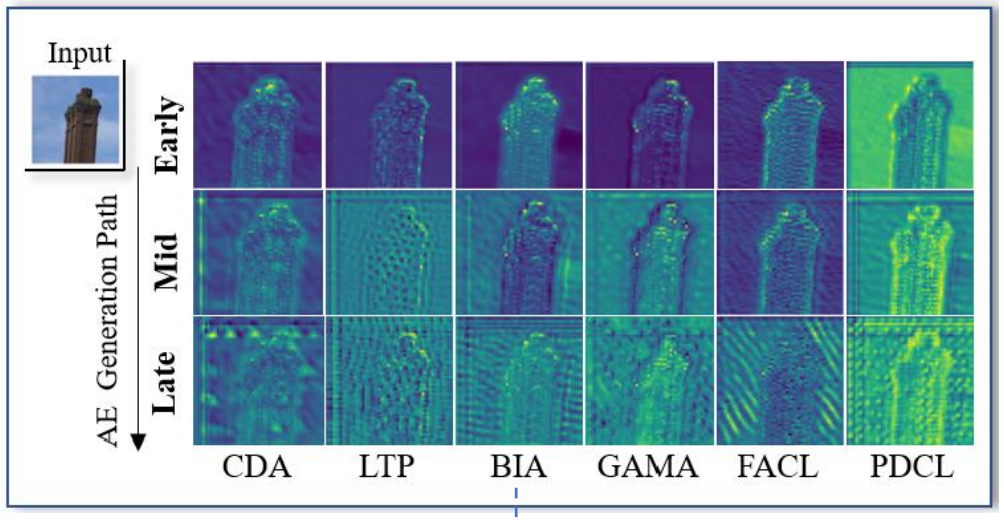
- **Generative Attacks** aim to train a perturbation generator that synthesizes adversarial examples in a single forward pass
  - **Fast and scalable black-box attack**
    - Amortizes attack cost by training a perturbation generator offline and producing adversarial examples in a single forward pass at test time
  - **Prior methods are surrogate-centric**
    - Exploits logit/feature/frequency/CLIP-based objectives defined on the surrogate model or perturbed image, rather than on the generator's own training dynamics
  - **Key limitation and our motivation**
    - Existing methods largely underexplore generator-internal feature synthesis
    - Our SCGA method explicitly focuses on generator feature-level semantic consistency, especially in the early blocks, to improve adversarial transfer without extra inference cost

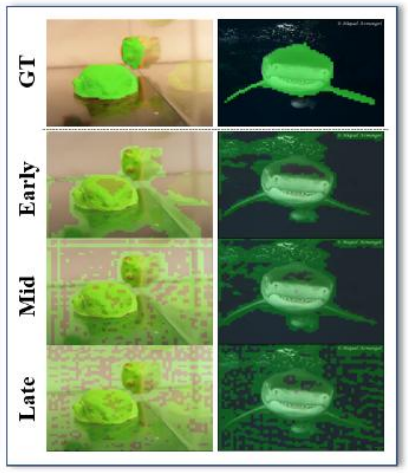
Our Focus

Input data aug.	Generator feature-level	Perturbed image-level	Surrogate mid-level layer feature-level	Surrogate output logit-level	GT label required?
-	-	-	-	✓	✓/✗
-	-	-	-	✓	✓/✗
-	-	-	✓	-	-
-	-	✓	✓	-	-
-	-	-	✓	-	✓
✓	-	-	✓	-	-
-	-	-	✓	-	✓
-	✓	-	-	-	-

Closer look into: Perturbation Generator

Delving into semantics in the intermediate layers of the generator





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**

Binary Mask Prediction on ImageNet-S (TPAMI 2022)

- ✓ **Insights on Internals of Perturbation Generator**
  - ✓ **Perturbations are synthesized progressively inside the generator**
    - Hence, transferability depends not only on the surrogate model (loss), but also on the generator's internal dynamics
  - ✓ **Early generator blocks retain object contours and coarse shape**
    - Whereas semantic recognizability degrades in mid to late blocks
  - ✓ **Lower semantic variability across blocks correlates with higher transferability**
    - Motivating early-block semantic anchoring in our SCGA method

Key Research Questions

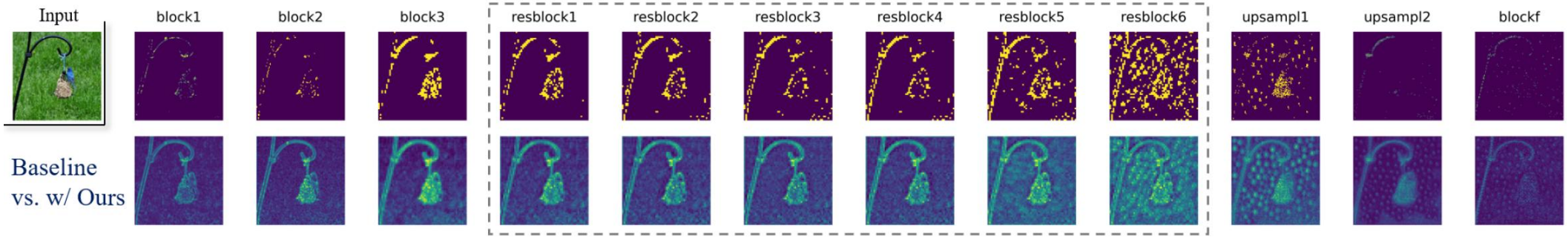
✓ Research Questions

1. When do semantic cues deteriorate?

- Object contours and coarse shapes are best preserved in early blocks, then progressively degrade in mid/late blocks

2. Which blocks semantically matter most for adversarial transferability?

- Early residual blocks matter most, as they shape the object-aligned scaffold that later blocks build upon



Block-wise perturbation generation progression

✓ Empirical Findings

- ✓ Lower cross-block semantic variability is associated with higher black-box transferability
- ✓ SCGA anchors the generator at the early intermediate blocks, so perturbations start on object-salient regions and then spread outward more effectively

Enforcing Semantic Consistency

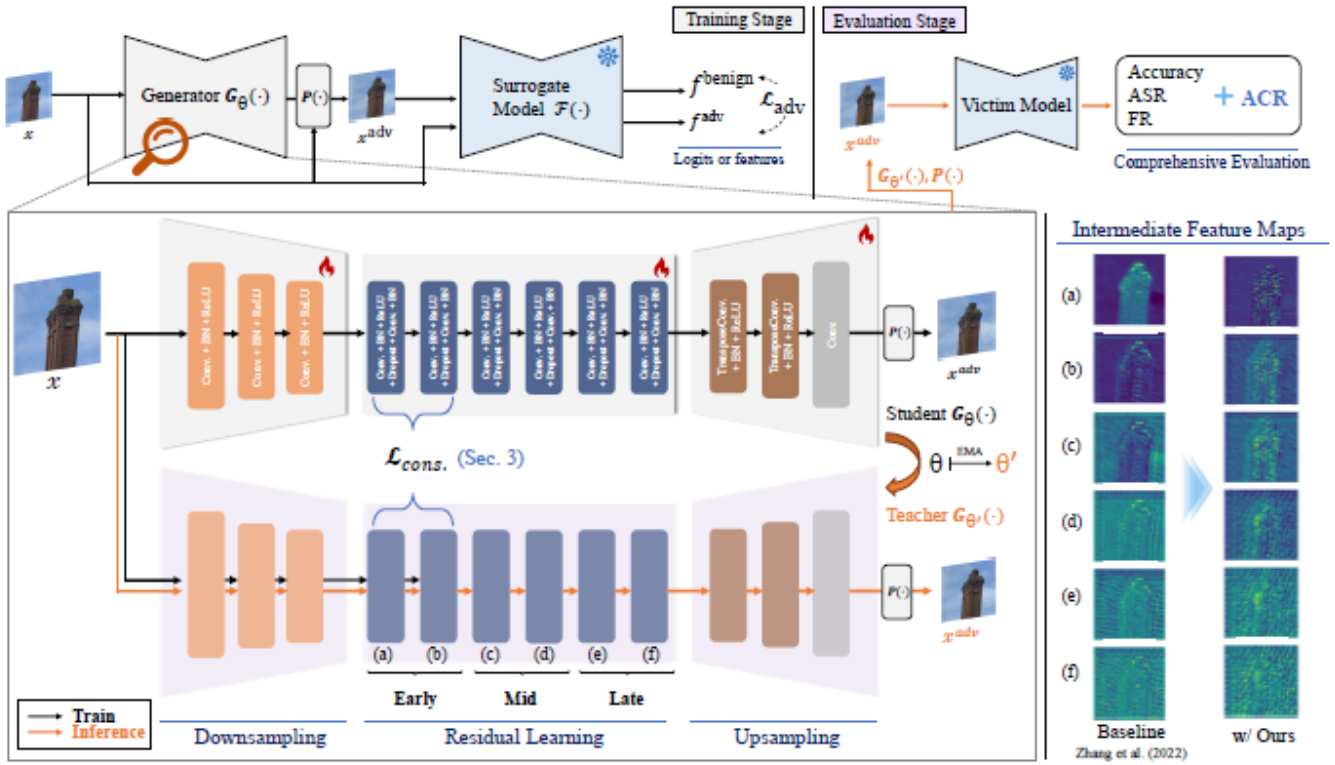
- **Preserving the underlying semantics within the generator**
  - **Stabilizing with a Mean Teacher**
    - Maintain a teacher generator as the EMA of the student as: $\theta'_t \leftarrow \eta\theta'_{t-1} + (1-\eta)\theta'_t$
    - Teacher provides temporally smoothed, less noisy feature references
  - **Early-block self-feature consistency**
    - Aligns the student features with the teacher's using a consistency loss to preserve object contours and coarse structure

$$\mathcal{L}_{\text{cons.}} = \sum_{\ell=1}^{L_{\text{early}}} \mathcal{W}_{\text{cons.}} \cdot \left[\tau - \frac{\langle \mathbf{g}_s^\ell, \mathbf{g}_t^\ell \rangle}{\|\mathbf{g}_s^\ell\|_2 \, \|\mathbf{g}_t^\ell\|_2} \right]_+$$

  - **Train with adversarial loss**

$$\mathcal{L}_{\text{adv}} = \cos(\mathcal{F}_k(x), \mathcal{F}_k(x^{adv}))$$

$$\boxed{\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda_{\text{cons.}} \cdot \mathcal{L}_{\text{cons.}}}$$



Overview of SCGA.

Accidental Correction Rate (ACR)

- ✓ **Limitation of conventional metrics**
  - Misses cases where the perturbation accidentally corrects an already misclassified sample
- ✓ **Novel Metric: ACR**
  - (Def) proportion of initially correct samples that become correct after attack

$$\mathrm{ACR} = \frac{|\{x \in I \mid f(x) \neq y,\ f(x+\delta) = y\}|}{|I|}$$

  - (Interpretation) ACR is a subset of fooling rate (FR), complementary to attack success rate (ASR), directly measuring attack reliability rather than mere disruption

GT Label	Benign Prediction	Adv. prediction	
Lorikeet	Ruffed Grouse ✗	Lorikeet ○	ACR
Ruffed Grouse	Ruffed Grouse ○	Lorikeet ✗	Conventional ASR, FR

Proposed ACR metric.

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 Setup

✓ **Training Protocol**
- Trained the perturbation generator on large-scale ImageNet-1K (1.2 M natural scene; 224×224) with $\varepsilon = 10/255$

✓ **Cross-setting evaluation**
- Model: ImageNet-1K trained image classification models (Acc.) (22 total, including CNNs, ViTs, Vision Mambas)
- Domain: CUB-200-2011, Stanford Cars, FGVC Aircraft (Acc.) (ResNet-50, SE-Net, SE-ResNet-101 models)
- Task: Semantic Segmentation (mIoU) (DeepLabV3+, SegFormer), Object Detection (mAP50) (Faster R-CNN, DETR)

✓ **Evaluation Metric**
- Top-1 Accuracy (Acc.)
- Attack Success Rate (ASR)
- Fooling Rate (FR)
- Accidental Correction Rate (ACR)

**for details, please refer to our paper*

Dataset Specifications.

Dataset	ImageNet-1K Russakovsky et al. (2015)	CUB-200-2011 Wah et al. (2011)	Stanford Cars Krause et al. (2013)	FGVC Aircraft Maji et al. (2013)
Train	1.2 M	5,994 (Not Used)	8,144 (Not Used)	6,667 (Not Used)
Val.	50,000	5,794	8,041	3,333
# Classes	1,000	200	196	100
Resolution	224×224	448×448	448×448	448×448

Results: Quantitative

✓ **Consistent cross-setting gains**

- SCGA improves transfer across model, domain, and task shifts for all baselines
- ACR decreases for every baseline, demonstrating fewer accidental corrections and thus more reliable attacks

✓ **Largest gains under harder shifts**

- More pronounced improvements for cross-domain transfer, especially on CDA and LTP (earlier works)
- Cross-task gains are also consistent, while GAMA, PDCL (more recent CLIP-based works) show smaller margins due to stronger pre-existing semantic guidance

Cross-setting black-box attack results.

Cross-	Model				Domain	Task	
Method	ΔAcc.↓	ΔASR↑	ΔFR↑	ΔACR↓	ΔAcc.↓	ΔSS (mIoU)↓	ΔOD (mAP50)↓
CDA w/ Ours	-6.89	+8.55	+7.49	-1.57	-15.12	-0.18	-0.71
LTP w/ Ours	-6.34	+8.47	+7.65	-0.70	-9.15	-0.86	-1.39
BIA w/ Ours	-1.04	+1.47	+1.28	-0.21	-3.96	-1.35	-0.20
GAMA w/ Ours	-1.12	+1.44	+1.28	-0.14	-2.47	-1.01	-0.16
FACL w/ Ours	-1.07	+1.35	+1.20	-0.28	-2.27	-0.88	-0.46
PDCL w/ Ours	-0.07	+0.09	+0.11	-0.03	-0.85	-1.04	-0.73

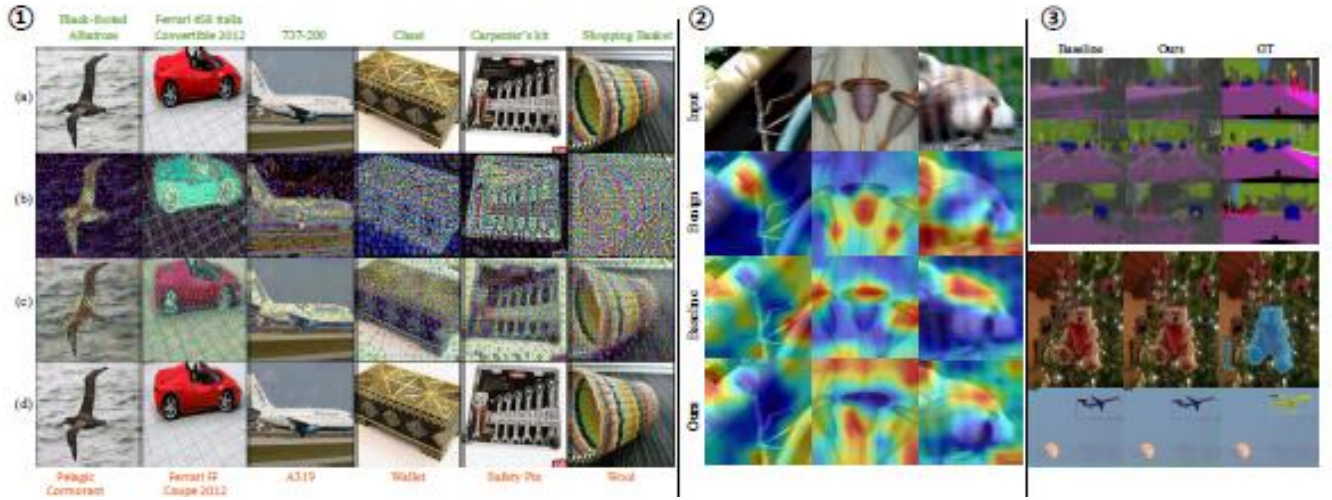
Results: Qualitative

✓ **Object-aligned perturbation synthesis**

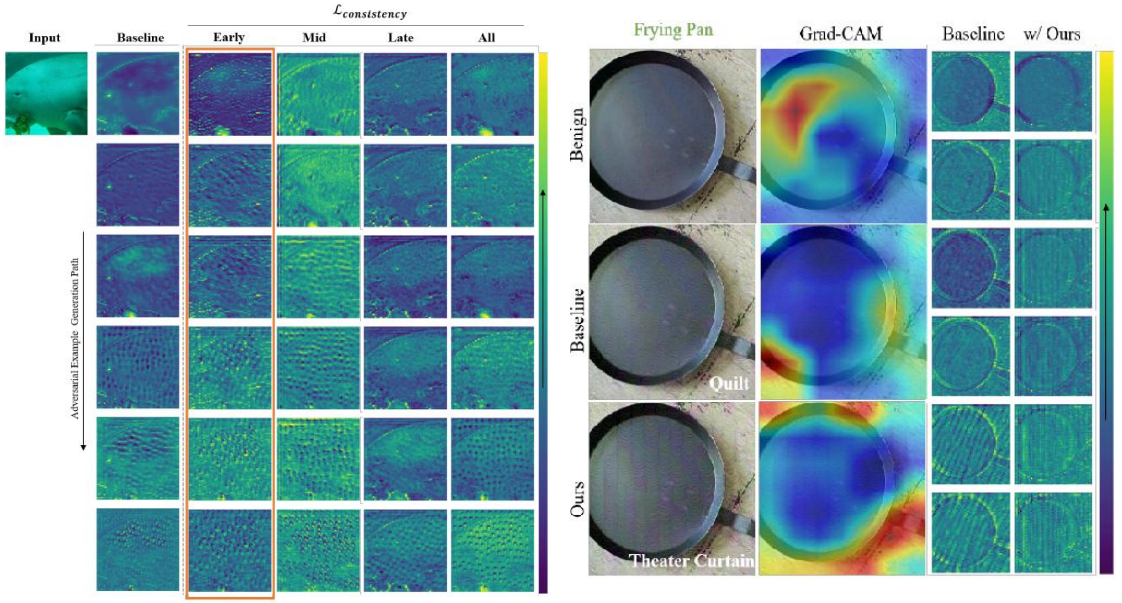
- SCGA places generated noise on semantic boundaries and salient object regions, instead of dispersing it arbitrarily
- Grad-CAM shows shifted and redistributed activations: stronger responses on semantically meaningful regions with less overly peaked behavior as in the baseline

✓ **Intermediate block-level activations**

- In the early blocks, perturbation differences concentrate on the foreground/object-centric regions
- In later blocks, the perturbation gradually spreads toward nearby background, **producing more transferable noise**



(a) Input, **(b)** Isolated generated perturbation, **(c)** unbounded adversarial example (AE), **(d)** bounded AE



Results: Using Different Surrogate

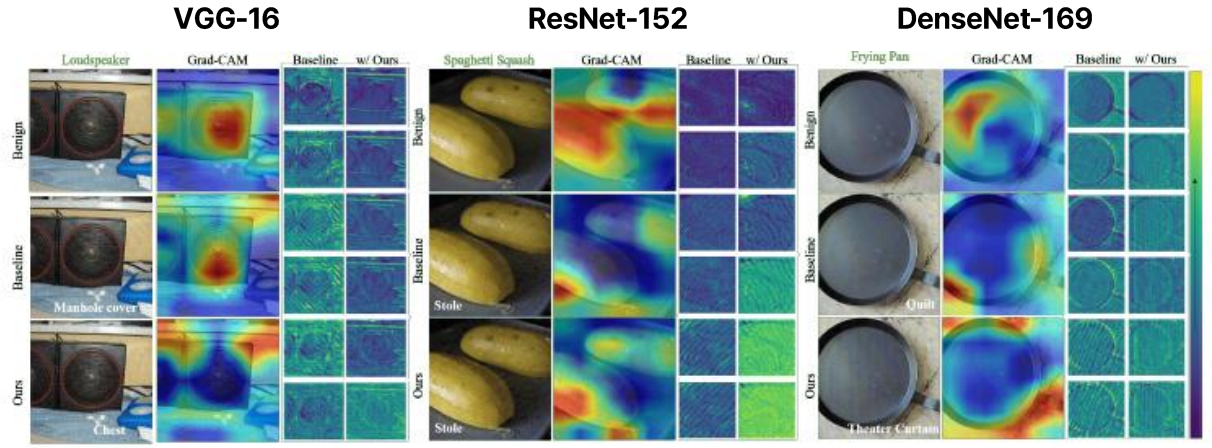
✓ **Robustness to surrogate choice**

- SCGA improves transferability even when trained with different surrogate models
  → the gain is not surrogate-specific ✓
- Notably, DenseNet-169 surrogate model gives slight boost in cross-model, while cross-domain improvements are substantial

Method	Cross-Model ΔAcc.↓	Cross-Model ΔASR↑	Cross-Model ΔFR↑	Cross-Model ΔACR↓	Cross-Domain ΔAcc.↓	Task ΔSS (mIoU)↓	Task ΔOD (mAP50)↓
BIA w/ Ours (VGG19)	-0.85	+1.10	+0.97	-0.14	-10.80	-1.64	+0.29
BIA w/ Ours (Res152)	-2.16	+2.74	+3.22	-0.74	-4.45	+2.50	+0.30
BIA w/ Ours (Dense169)	-2.90	+3.69	+3.22	-0.74	-7.94	-1.33	-0.56

✓ **Consistent improvements across surrogates**

- Grad-CAMs reveal broader high-sensitivity regions than the baseline, rather than focusing only on the dominant object region
- Perturbations consistently align with object edges and contours
  → SCGA captures shared, model-agnostic semantics ✓



✓ **Stronger transfer with preserved perceptual quality**

- SCGA improves cross-model/domain/task performance while maintaining the perceptual quality
  → no degradation in perceptual quality

Method	Cross-setting (Avg.)				Perceptual Quality		
	Domain (Acc.)	Model (Acc.)	Task (SS; mIoU)	Task (OD; mAP50)	PSNR ↑	SSIM ↑	LPIPS ↓
CDA	69.94	50.27	22.90	29.54	29.11	0.78	0.43
w/ Ours	**54.82**	**43.38**	**22.71**	**28.83**	29.17 (+0.06)	0.78 (–)	0.43 (–)
LTP	49.91	48.33	25.34	25.90	29.11	0.76	0.47
w/ Ours	**40.76**	**41.99**	**24.48**	**24.52**	29.26 (+0.15)	0.77 (+0.01)	0.49 (+0.02)
BIA	51.07	45.17	24.75	24.72	28.08	0.75	0.49
w/ Ours	**47.10**	**44.13**	**23.40**	**24.52**	28.76 (+0.68)	0.75 (–)	0.49 (–)
GAMA	48.56	44.58	25.82	24.36	28.62	0.74	0.49
w/ Ours	**46.09**	**43.46**	**24.81**	**24.20**	28.69 (+0.07)	0.74 (–)	0.49 (–)
FACL	44.05	42.00	25.08	24.43	28.61	0.74	0.49
w/ Ours	**41.78**	**40.92**	**24.20**	**23.97**	28.67 (+0.05)	0.74 (–)	0.49 (–)
PDCL	43.91	42.84	25.24	24.93	28.68	0.74	0.48
w/ Ours	**43.06**	**42.77**	**24.20**	**24.20**	28.70 (+0.02)	0.74 (–)	0.49 (–)

✓ **Semantic frequency shift inside the generator**

- SCGA generally increases low-band energy (↑) and reduces high-band energy (↓) across early/mid/late stages, supporting better preservation of coarse semantic structure
  → SCGA suppresses surrogate-specific artifacts and propagates a more stable semantic scaffold through later blocks

	Band	Early	Mid	Late
CDA	Low (↑)	0.82→**0.91**	0.75→**0.97**	0.77→**0.96**
→w/ Ours	High (↓)	0.18→**0.09**	0.25→**0.03**	0.23→**0.04**
LTP	Low (↑)	0.73→0.72	0.78→**0.79**	0.95→0.75
→w/ Ours	High (↓)	0.27→0.28	0.22→**0.21**	0.05→0.25
BIA	Low (↑)	0.56→0.56	0.53→**0.54**	0.53→**0.58**
→w/ Ours	High (↓)	0.44→0.44	0.47→**0.45**	0.47→**0.42**
GAMA	Low (↑)	0.57→**0.79**	0.54→**0.60**	0.56→**0.59**
→w/ Ours	High (↓)	0.43→**0.21**	0.46→**0.40**	0.44→**0.41**
FACL	Low (↑)	0.57→**0.73**	0.52→**0.61**	0.54→**0.59**
→w/ Ours	High (↓)	0.43→**0.27**	0.48→**0.39**	0.46→**0.45**
PDCL	Low (↑)	0.54→**0.62**	0.51→**0.59**	0.58→**0.59**
→w/ Ours	High (↓)	0.46→**0.38**	0.49→**0.41**	0.42→**0.41**

Computational Cost

✓ **Maintains Compute Efficiency**

- No inference-time overhead
- Small training overhead
- The extra cost primarily stems from teacher-generator forward pass during training

Method	Train time (hh:mm)	Peak memory (MB)	GPU type	Train batch size
Baseline	5:00	1,384.62	NVIDIA RTX A6000 (1×)	48
w/ Ours	5:40	1,442.23		

Batch size=1

Method	Student fwd (ms)	Teacher fwd (ms)	Backward (ms)	Total (ms)	Backward cost (GFLOPs)	Backward CUDA time (ms)
Baseline	7.1	–	25.6	32.7	0.0012	19.714
w/ Ours	6.8	3.9	28.4	39.1	0.0044	20.192

Robustness to Defenses

✓ **Robustness to defense mechanisms**

- SCGA improves attack effectiveness against adversarially trained models, standard defenses (JPEG, BDR, R&P), and input transformation (rotation, smoothing, TVM, pixel deflection) defenses
  → SCGA generally further induces Acc./ACR ↓ and ASR/FR ↑

Method	Metric	Adv.IncV3	Adv.ViT	Adv.ConvNeXt	JPEG	BDR	R&P	Avg.
Benign	Acc. (%) ↓	76.33	48.82	58.44	74.68	74.68	76.58	68.26
Baseline Zhang et al. (2022b)	Acc. (%) ↓	68.54	45.64	53.88	63.49	47.82	44.78	54.03
	ASR (%) ↑	14.95	11.72	10.26	20.24	40.76	44.59	23.75
	FR (%) ↑	24.02	25.48	19.40	28.09	48.06	51.60	32.78
	ACR (%) ↓	15.30	4.96	3.46	11.45	11.30	10.56	9.51
w/ Ours	Acc. (%) ↓	67.92	45.33	53.62	60.83	44.07	39.01	**51.80**
	ASR (%) ↑	15.75	11.95	10.65	23.74	45.37	51.63	**26.52**
	FR (%) ↑	24.83	25.31	19.60	31.61	52.22	57.86	**35.28**
	ACR (%) ↓	15.23	4.57	3.38	11.48	10.29	9.08	**9.01**

	AT + Common Defenses (Adv.IncV3, Adv.ViT, Adv.ConvNeXt, JPEG, BDR, R&P)			
Method	ΔAcc.↓	ΔASR↑	ΔFR↑	ΔACR↓
BIA	-2.23	+2.77	+2.50	-0.50
GAMA	-0.73	+0.91	+0.72	-0.14
FACL	-1.59	+1.96	+1.62	-0.40
PDCL	-0.40	+0.56	+0.50	+0.02

Additional Input Processing (Rot. 30/50/70/90, Gaussian/Median/Mean, TVM, PD)				Purification (NRP, NRP-ResNet)			
ΔAcc.↓	ΔASR↑	ΔFR↑	ΔACR↓	ΔAcc.↓	ΔASR↑	ΔFR↑	ΔACR↓
-3.03	+3.80	+3.45	-0.56	-0.26	+0.39	+0.56	+0.19
-1.36	+1.70	+1.48	-0.26	-0.31	+0.55	+0.80	+0.41
-1.82	+2.28	+2.02	-0.36	+0.02	-0.04	-0.14	-0.07
-0.54	+0.89	+0.79	-0.07	+0.17	-0.28	-0.37	-0.17

Application to Targeted Attack

✓ **Extension to targeted black-box attacks**

- SCGA consistency improves target success rate (TSR) when applied to strong targeted generators such as CGNC and M3D
  → Consistent gains across victims

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**
ResNet152											
M3D Zhao et al. (2023a)	50.73	62.76	63.15	69.53	60.96	57.82	37.28	73.87	37.82	11.29	52.52
w/ Ours	60.35	58.80	68.61	79.73	73.64	65.47	56.12	70.29	52.77	39.63	**62.54**
WRN-50-2											
M3D Zhao et al. (2023a)	64.55	68.95	73.51	69.53	73.04	66.53	46.85	84.39	45.17	13.15	60.57
w/ Ours	72.41	69.22	77.94	79.61	82.97	69.83	66.16	76.32	58.20	41.01	**69.37**

	Victim model							Avg.
Method	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**

Key Takeaway

- ✓ **Generator internals are a key source of transferability**
  - Adversarial transfer is shaped not only be surrogate objectives, but also by how perturbations evolve inside the perturbation generator

- ✓ **Early semantic anchoring improves both attack strength and reliability**
  - SCGA preserves object-aligned structure in early blocks, yielding higher ASR/FR and lower ACR

- ✓ **SCGA as a broadly effective plug-and-play method**
  - Works as a training-only add-on with no inference overhead
  - Generalizes across baselines, domains, models, tasks (as well as defenses and targeted attacks)

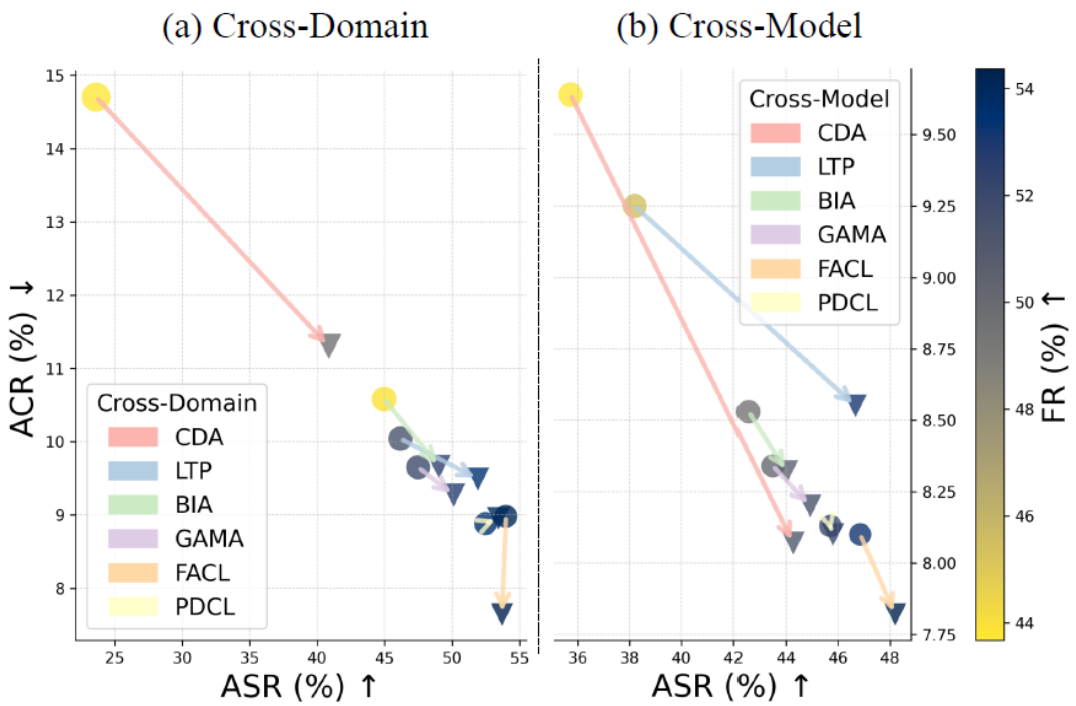
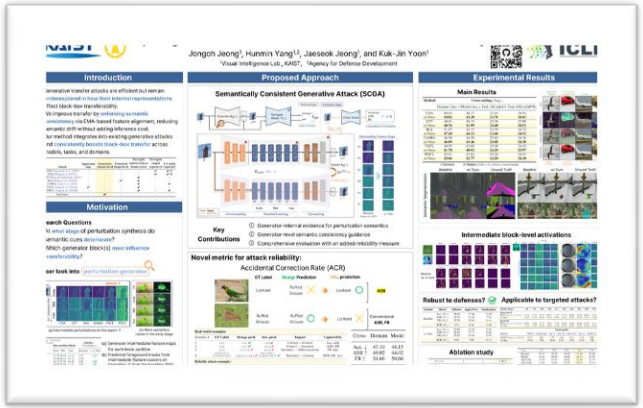


Figure: Our semantically consistent generative attack leverages generator intermediate features to craft adversarial examples that improve transferability over the baselines, (baseline ● → ours ▼), across domains in (a) and across models in (b).

Improving Black-Box Generative Attacks via Generator Semantic Consistency



Project Page

Thank you!

Poster Session:
Fri, Apr 24, 2026 | 3:15 PM – 5:45 PM