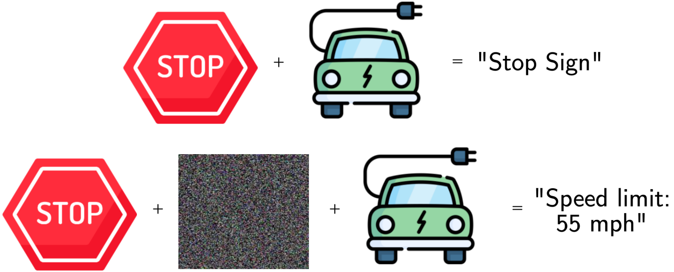
GAMA: Generative Adversarial Multi-Object Scene Attacks

Abhishek Aich[1], Calvin-Khang Ta[1], Akash Gupta, Chengyu Song, Srikanth V. Krishnamurthy, M. Salman Asif, Amit K. Roy-Chowdhury

[1] joint first authors

Adversarial Attacks

- Bad actors/attackers are always looking to break systems
  - ⇝ self-driving cars, face-identification systems, etc.



Adversarial Attacks

◆ Attackers are evolving $\cdots$ and so are their attacking tools!

$\rightsquigarrow$ Past $\sim$5 years, focus on generative adversarial attacks

$\rightsquigarrow$ Generative Attacks use surrogate models[1,2,3,4]

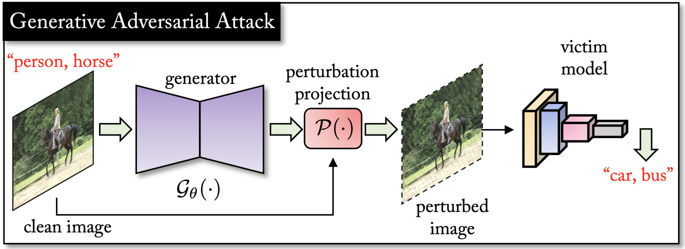
[1] Omid Poursaeed et al. "Generative Adversarial Perturbations". *CVPR*. 2018.

[2] Muzammal Naseer et al. "Cross-Domain Transferability of Adversarial Perturbations". *NeurIPS* (2019).

[3] Mathieu Salzmann et al. "Learning Transferable Adversarial Perturbations". *NeurIPS* (2021).

[4] Qilong Zhang et al. "Beyond ImageNet Attack: Towards Crafting Adversarial Examples for Black-box Domains". *ICLR*. 2022.

Adversarial Attacks



- Generative attacks are characterized by
  - High transferability of perturbations
  - Perturb large number of images with one forward pass

Problem Statement

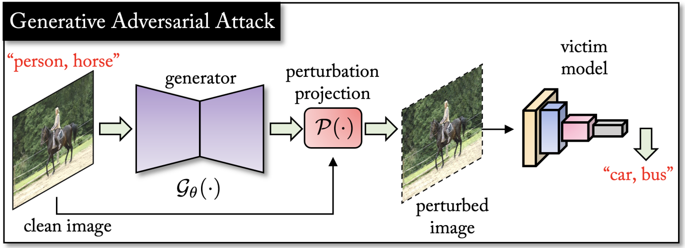
- Prior works only focused on perturbing scenes with one object
  - ⤳ e.g. datasets like ImageNet, CIFAR100
- But natural/real-world scenes contain multiple objects
  - ⤳ e.g. datasets like Pascal-VOC, MS-COCO

single-object scenes

multi-object scenes

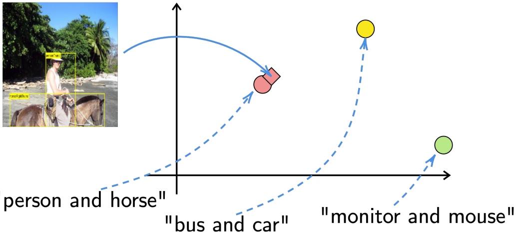
Problem Statement

Design a generative attack for multi-object scenes which crafts imperceptible perturbations to fool multi-label classifiers



Vision-Language models for Attacks (!)

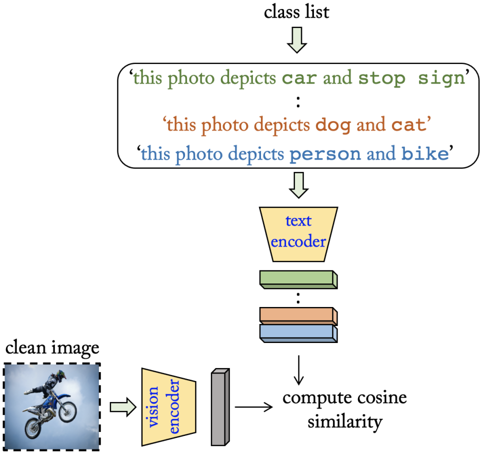
- “Contrastive Language–Image Pre-training” framework or CLIP[5]
  - ⇝ pre-trained on ∼400 million images, open-sourced
  - ⇝ provides generalized image features
  - ⇝ (most importantly), allows language-image alignment property



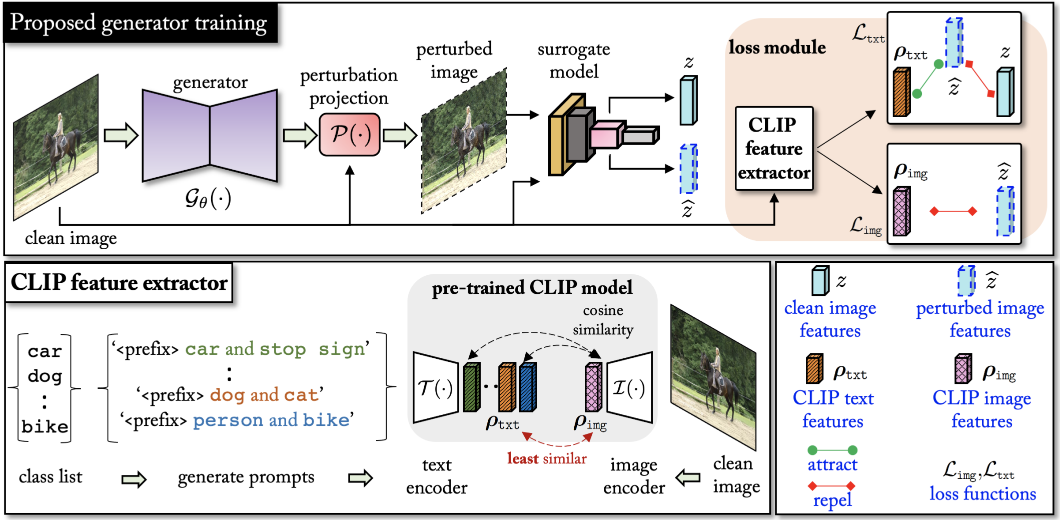
[5] Alec Radford et al. “Learning transferable visual models from natural language supervision”. *ICML*. 2021.

Vision-Language models for Attacks (!)

- CLIP can be "exploited" by the attacker
- Natural scenes have co-occurring objects
- These contextual relationships can be easily encoded in language
  - ⤳ e.g. "person" and "horse" $\rightarrow$ "a photo depicts person and horse"



Vision-Language models for Attacks (!)



Attack scenarios

- $f(\cdot)$ is the surrogate model trained on distribution $\mathcal{D}$
- $g(\cdot)$ is the victim model trained on distribution $\mathcal{D}_t$
  - $\rightsquigarrow$ *Scenario 1*: an attack termed *white-box* if $f(\cdot) = g(\cdot)$ and $\mathcal{D} = \mathcal{D}_t$
  - $\rightsquigarrow$ *Scenario 2*: an attack termed *black-box* if either $f(\cdot) \neq g(\cdot)$ or $\mathcal{D} \neq \mathcal{D}_t$

Same-Distribution Attack Results

◆ GAMA creates strong perturbations under both white-box and black-box attacks

Table 1: Pascal-VOC → Pascal-VOC (white-box attacks)

$f(\cdot)$	Method	VGG16	VGG19	Res50	Res152	Den169	Den121	Average
	No Attack	82.51	83.18	80.52	83.12	83.74	83.07	82.69
VGG19	GAP [1]	19.64	16.60	72.95	76.24	68.79	66.50	53.45
	CDA [2]	26.16	20.52	61.40	65.67	70.33	62.67	51.12
	TAP [3]	24.77	19.26	66.95	66.95	68.65	64.51	51.84
	BIA [4]	12.53	14.00	64.24	69.07	69.44	64.71	48.99
	GAMA	**6.11**	**5.89**	**41.17**	**45.57**	**53.11**	**44.58**	**32.73**
Res152	GAP [1]	56.93	56.20	65.58	72.26	75.22	69.54	65.95
	CDA [2]	41.07	47.60	53.84	47.22	67.50	59.65	52.81
	TAP [3]	52.92	58.24	56.52	53.61	71.55	64.56	59.56
	BIA [4]	45.34	49.74	51.98	50.27	67.75	61.05	54.35
	GAMA	**33.42**	**39.42**	**32.39**	**20.46**	**49.76**	**49.54**	**37.49**

(hamming scores in %, lower is better)

Different-Distribution Attack Results

◆ GAMA shows strong transferability of perturbations for stricter black-box attacks

Table 2: Pascal-VOC → ImageNet

$f(\cdot)$	Method	VGG16	VGG19	Res50	Res152	Den121	Den169	Average
	No Attack	70.15	70.94	74.60	77.34	74.22	75.74	73.83
VGG19	GAP [1]	24.44	21.64	63.65	67.84	63.09	65.47	51.02
	CDA [2]	13.83	11.99	47.32	53.92	46.81	52.24	37.68
	TAP [3]	06.70	07.28	50.94	57.36	47.68	53.43	37.23
	BIA [4]	04.20	04.73	48.63	57.65	45.94	53.37	35.75
	GAMA	**03.07**	**03.41**	**22.32**	**34.04**	**24.51**	**30.35**	**19.61**
Res152	GAP [1]	34.04	34.67	52.85	61.61	58.09	59.24	50.08
	CDA [2]	29.33	34.88	44.28	46.05	46.91	51.62	42.17
	TAP [3]	33.25	37.53	41.18	42.14	50.96	56.45	43.58
	BIA [4]	22.82	27.44	34.66	36.74	45.48	51.26	36.40
	GAMA	**16.43**	**17.02**	**21.93**	**17.07**	**31.63**	**30.57**	**22.44**

(hamming scores in %, lower is better)

Classifier-to-Detector Attack Results

- GAMA crafts better perturbations even for extreme black-box attacks

Table 3: Pascal-VOC → MS-COCO Object Detection task

$f(\cdot)$	Method	FRCN	RNet	DETR	D^2ETR	Average
	No Attack	0.582	0.554	0.607	0.633	0.594
VGG19	GAP [1]	0.424	0.404	0.360	0.410	0.399
	CDA [2]	0.276	0.250	0.208	0.244	0.244
	TAP [3]	0.384	0.340	0.275	0.320	0.329
	BIA [4]	0.347	0.318	0.253	0.281	0.299
	GAMA	**0.234**	**0.207**	**0.117**	**0.122**	**0.170**
Res152	GAP [1]	0.389	0.362	0.363	0.408	0.380
	CDA [2]	0.305	0.274	0.256	0.281	0.279
	TAP [3]	0.400	0.348	0.288	0.350	0.346
	BIA [4]	0.321	0.275	0.205	0.256	0.264
	GAMA	**0.172**	**0.138**	**0.080**	**0.095**	**0.121**

(`bbox_mAP_50` values, lower is better)

Adversarial examples

top row: clean images, **bottom row**: perturbed images,
text on each image: victim classifier predictions

Thank You!

- **Acknowledgement** → This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00112090096.
- **Paper ID: 130** → GAMA: Generative Adversarial Multi-Object Scene Attacks

(Project page)