Prompt-Driven Contrastive Learning for Transferable Adversarial Attacks



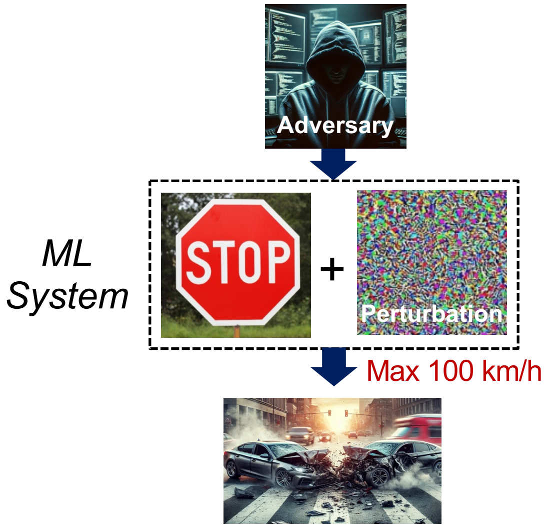
Hunmin Yang[1,2] Jongoh Jeong[1] Kuk-Jin Yoon[1]

[1] KAIST [2] ADD

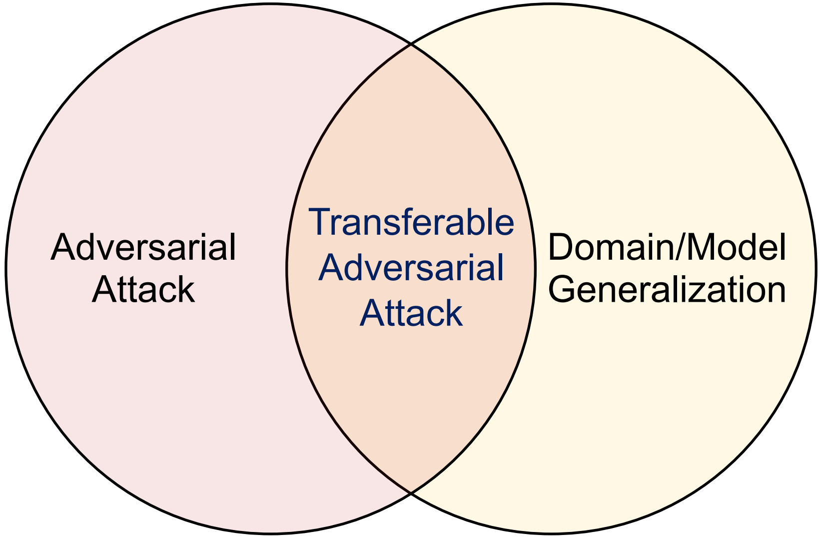
EUROPEAN CONFERENCE ON COMPUTER VISION

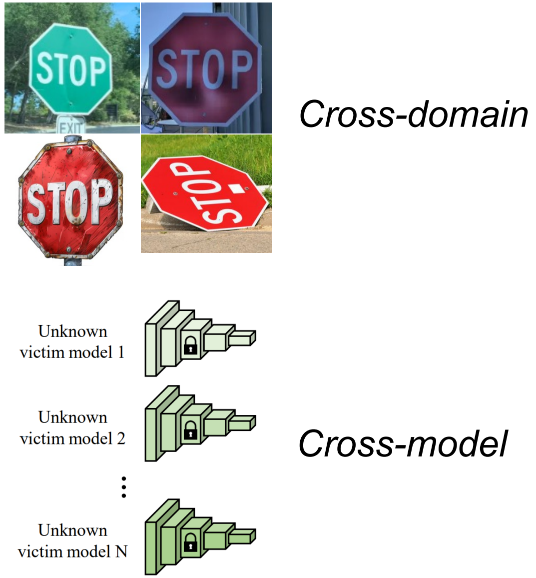
Introduction

- Transferable Adversarial Attacks
  - Crafting adversarial perturbations that are transferable to unknown domains and models.
- Significance
  - Security concerns: Identifying vulnerabilities in ML systems (e.g., autonomous driving)
  - Robustness testing: Serving as a benchmark for evaluating the ML robustness.



Copyright : DALL-E 3

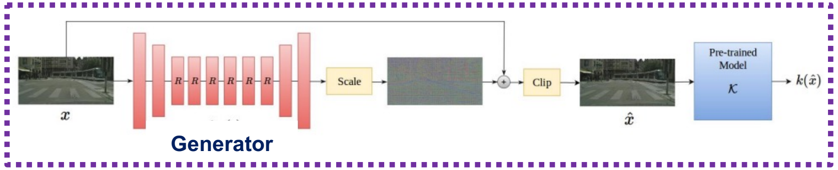




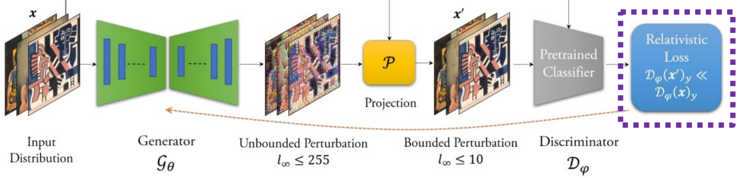
Introduction

- Recent progress in generator-based transferable attacks (1/2)

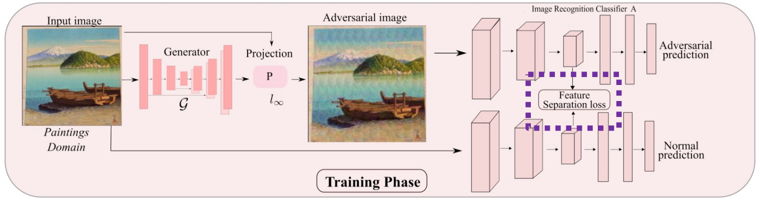
GAP [CVPR'18]



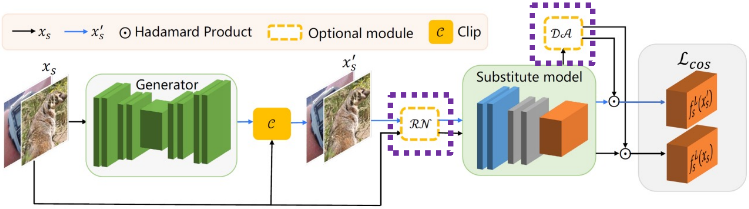
CDA [NeurIPS'19]



LTP [NeurIPS'21]



BIA [ICLR'22]

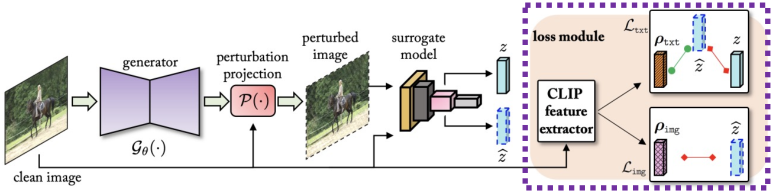


- Generator-based attack framework
- Cross-domain attack via relativistic CE loss
- Using mid-layer features of surrogate model
- RN (random data normalization)
  + DA (domain-agnostic feature attention)

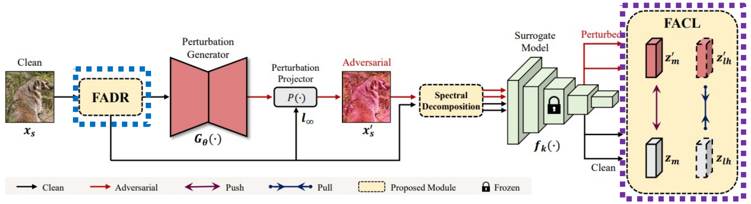
Introduction

- Recent progress in generator-based transferable attacks (2/2)

GAMA [NeurIPS'22]



FACL-Attack [AAAI'24]



- CLIP model for multi-object scene attacks
- FADR (domain-specific data randomization) + FACL (domain-agnostic feature contrast)

- Key insights and our hypothesis

Identifying generalizable representations across diverse domains and models	→	Training a robust perturbation generator that is both domain- and model-agnostic

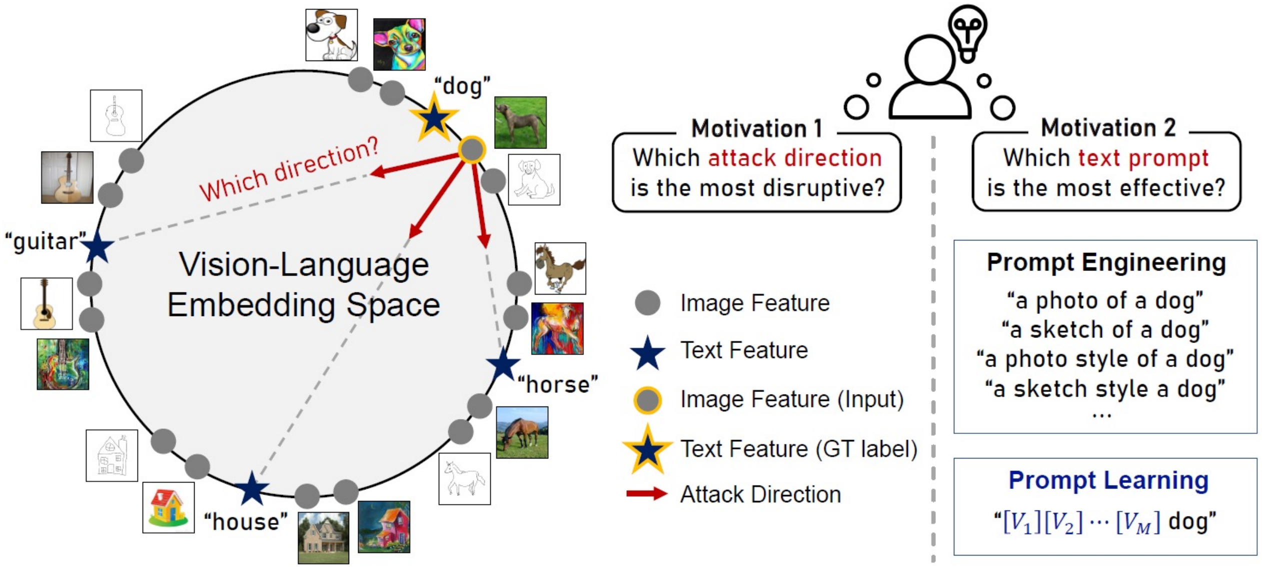
"Foundation model guidance could lead to more effective attacks."

Motivation & Goal

- A new comer: CLIP
  - A vision-language foundation model with highly generalizable representations
    - In a joint vision-language space, <u>a single text can represent various images from diverse domains.</u>

      Motivation 1
    - <u>The type of text prompt greatly affects the effectiveness</u>.

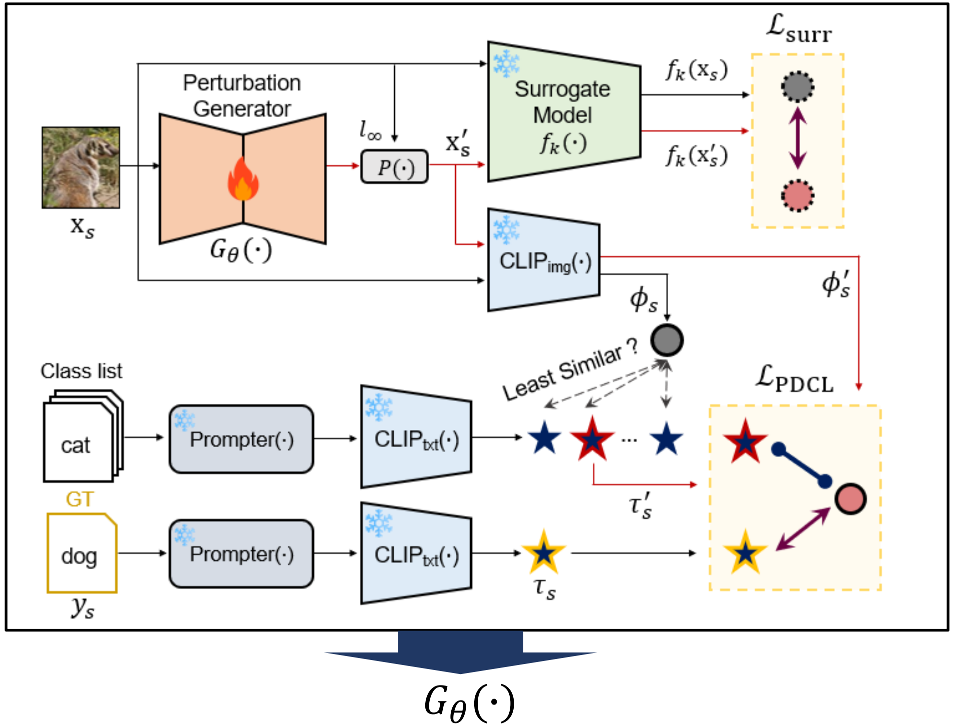
      Motivation 2

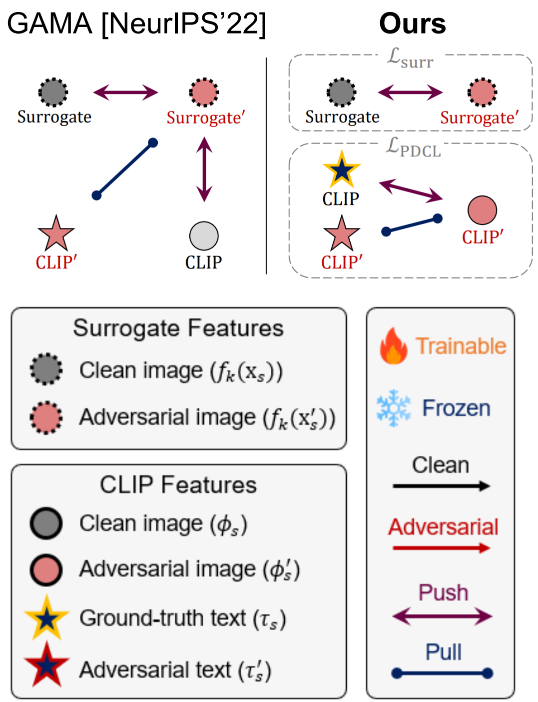


► Goal 1: Prompt-driven attack guidance ► Goal 2: Robust prompting via learning

Proposed Method

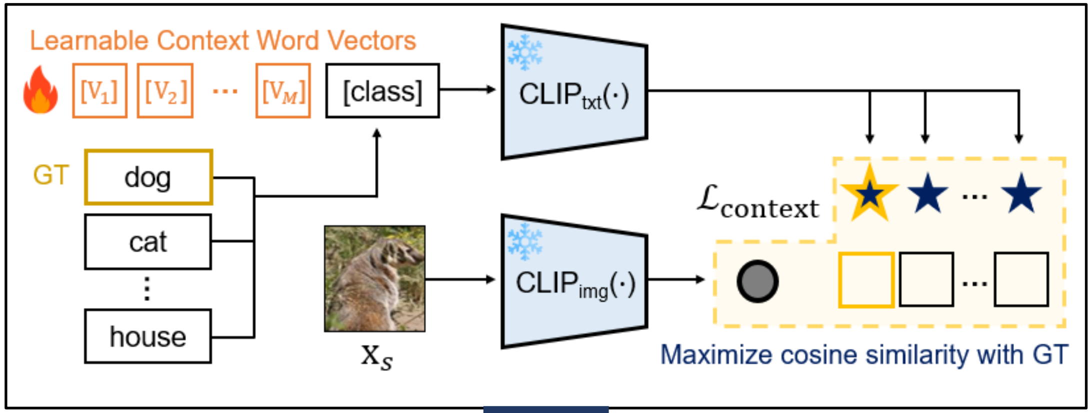
- [Method 1] Prompt-driven attack guidance
  - Leveraging prototypical text features, our prompt-driven contrastive loss $\mathcal{L}_{\mathrm{PDCL}}$ improves the robustness of the perturbation generator $G_\theta(\cdot)$ to diverse input images.
    - Our loss separately deals with feature spaces of surrogate model and CLIP model.



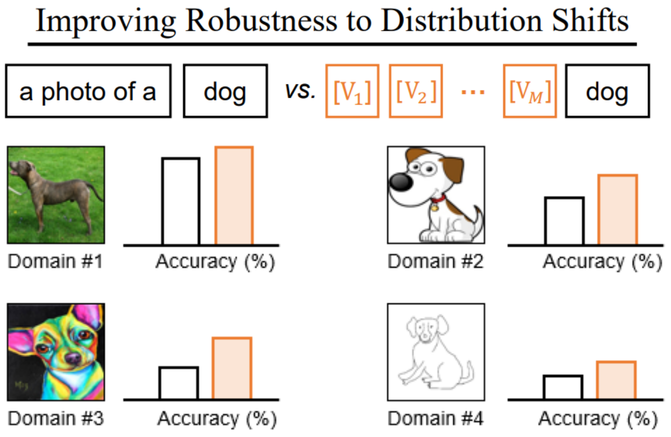


Proposed Method

- [Method 2] Robust prompting via learning
  - For effective prompt-driven feature guidance, we pre-train learnable context word vectors using $\mathcal{L}_{\text{context}}$ to produce more generalizable text features [Zhou et al., 2022].
    - With the frozen CLIP model, we train only the learnable context word vectors of Prompter(·).



Prompter(·) = $[V_1]\ [V_2] \cdots [V_M]$ [·]



Distribution-shifted Variants

Method	ImageNet-1K	-V2	-Sketch	-A	-R
Zero-shot CLIP [41]	66.7	60.9	46.1	47.8	74.0
w/ Prompter(·)	**71.9**	**64.2**	**46.3**	**48.9**	**74.6**

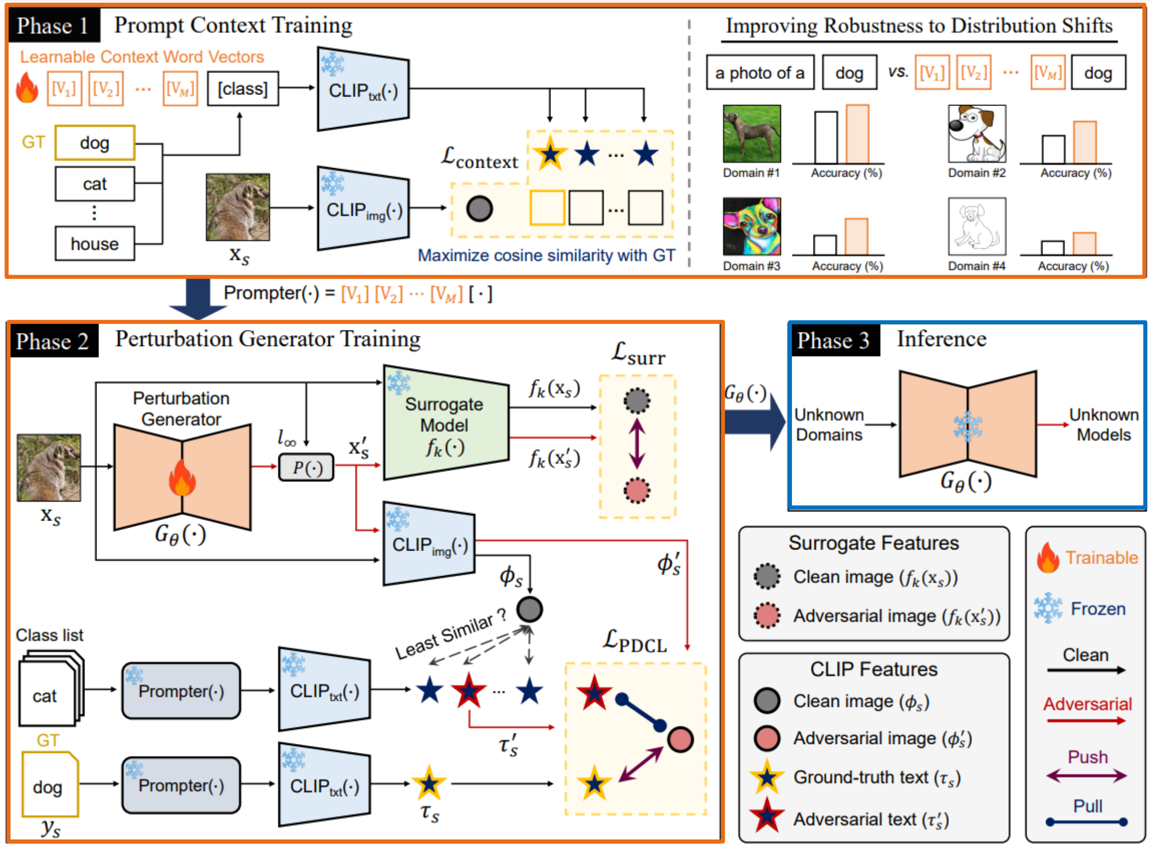
[Zhou et al., 2022] Learning to Prompt for Vision-Language Models, IJCV 2022

Proposed Method

- Overall framework
  - Two training phases (Phase 1 & Phase 2) and an inference phase (Phase 3)

[Method 2]
Robust prompting via learning

[Method 1]
Prompt-driven attack guidance



Inference

Experimental Results

Cross-Domain Transferability

Unknown
domain 1

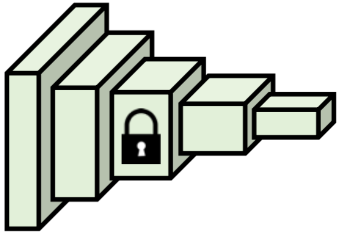
Unknown
domain 2

⋮

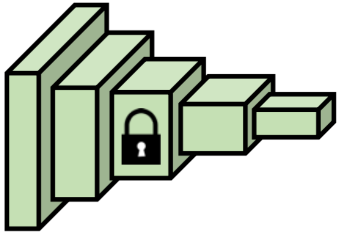
Unknown
domain N

Cross-Model Transferability

Unknown
victim model 1

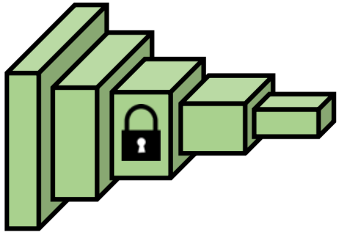


Unknown
victim model 2



⋮

Unknown
victim model N



Experimental Results

- Cross-domain attack transferability

Method	CUB-200-2011			Stanford Cars			FGVC Aircraft			AVG.
	Res-50	SENet154	SE-Res101	Res-50	SENet154	SE-Res101	Res-50	SENet154	SE-Res101	
Clean	87.33	86.81	86.59	94.25	93.35	92.96	92.14	92.05	91.84	90.81
GAP [40]	68.85	74.11	72.73	85.64	84.34	87.84	81.40	81.88	76.90	79.30
CDA [35]	69.69	62.51	71.00	75.94	72.45	84.64	71.53	58.33	63.39	69.94
LTP [34]	<u>30.86</u>	<u>52.50</u>	62.86	34.54	<u>65.53</u>	73.88	<u>15.90</u>	60.37	52.75	49.91
BIA [60]	32.74	52.99	58.04	39.61	69.90	70.17	28.92	60.31	<u>46.92</u>	51.07
GAMA [2]	34.47	54.02	<u>57.66</u>	<u>30.16</u>	69.80	<u>63.82</u>	25.29	<u>58.42</u>	**43.41**	<u>48.56</u>
Ours	**22.97**	**49.19**	**54.92**	**22.58**	**64.95**	**63.70**	**15.81**	**53.83**	47.25	**43.91**

- **Domain**: ImageNet-1K (Source Domain) → CUB, CAR, AIR (Target Domain)
- **Model**: VGG-16 (Surrogate Model) → Various Models (Victim Models)

Experimental Results

- Cross-model attack transferability

Method	Res-50	Res-152	Dense-121	Dense-169	Inc-v3	MNasNet	ViT-B/16	ViT-L/16	AVG.
Clean	74.61	77.34	74.22	75.75	76.19	66.49	79.56	80.86	75.63
GAP [40]	57.87	65.50	57.94	61.37	63.30	42.47	72.89	76.69	54.34
CDA [35]	36.27	51.05	38.89	42.67	54.02	33.10	68.73	74.22	53.24
LTP [34]	<u>21.70</u>	<u>39.88</u>	<u>23.42</u>	**25.46**	41.27	45.28	72.44	76.75	43.28
BIA [60]	25.36	42.98	26.97	32.35	41.20	34.31	<u>67.05</u>	73.23	42.93
GAMA [2]	24.82	43.22	24.84	30.81	<u>35.10</u>	**27.96**	67.33	<u>73.16</u>	<u>40.91</u>
Ours	**20.87**	**38.62**	**21.26**	<u>29.01</u>	**32.99**	<u>28.00</u>	**65.53**	**72.52**	**38.60**

- **Domain** : ImageNet-1K (Source Domain = Target Domain)
- **Model** : VGG-16 (Surrogate Model) → Various Models (Victim Models)

Experimental Results

- Ablation study on our proposed losses

Method	$\mathcal{L}_{\text{surr}}$	$\mathcal{L}_{\text{GAMA}}$	$\mathcal{L}_{\text{PDCL}}$	$\mathcal{L}_{\text{context}}$	Cross-Domain	Cross-Model
Clean	–	–	–	–	90.85	75.63
BIA [60]	✓	–	–	–	51.07	42.93
GAMA [2]	✓	✓	–	–	48.56	40.91
Ours†	✓	–	✓	–	<u>46.69</u>	<u>40.35</u>
Ours	✓	–	✓	✓	**43.91**	**38.60**

- Our proposed $\mathcal{L}_{\text{PDCL}}$ achieves SoTA even w/o prompt learning of $\mathcal{L}_{\text{context}}$.

Experimental Results

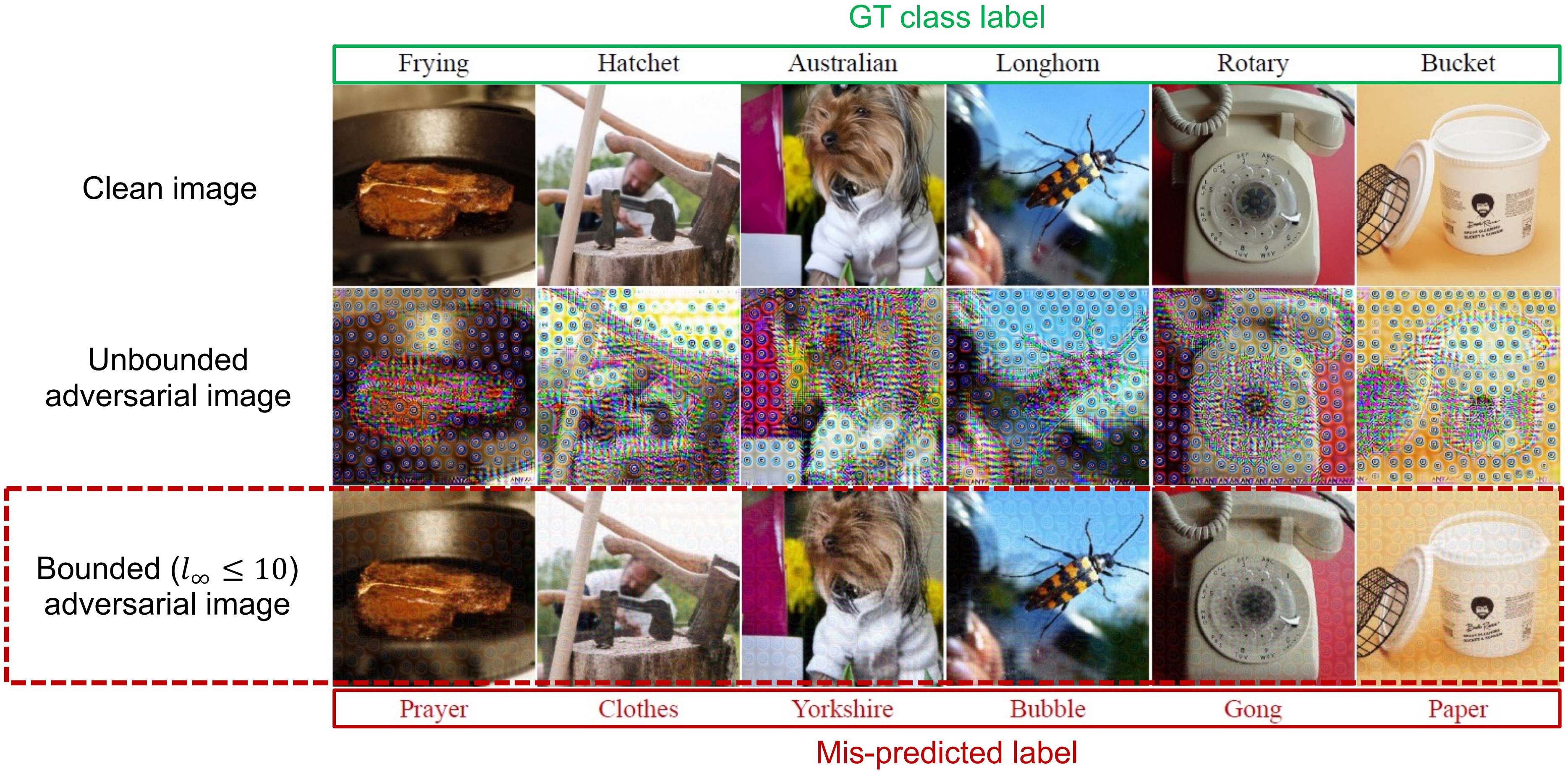
- Effect of learnable context words

Type	# of words	Text Prompt	Accuracy (↓)
Heuristic	$M = 4$	"a photo of a [class]"	46.69
		"a sketch of a [class]"	47.02
	$M = 5$	"a photo style of a [class]"	46.14
		"a sketch style of a [class]"	47.70
		"a $[\mathbf{V}_{\text{rand}}]$ style of a [class]"	47.81
Learnable	$M = 4$	"$[\mathbf{V}_1][\mathbf{V}_2][\mathbf{V}_3][\mathbf{V}_4]$ [class]"	<u>45.44</u>
	$M = 16$	"$[\mathbf{V}_1][\mathbf{V}_2]\cdots[\mathbf{V}_{16}]$ [class]"	**43.91**

- Prompt learning is more effective than engineering.

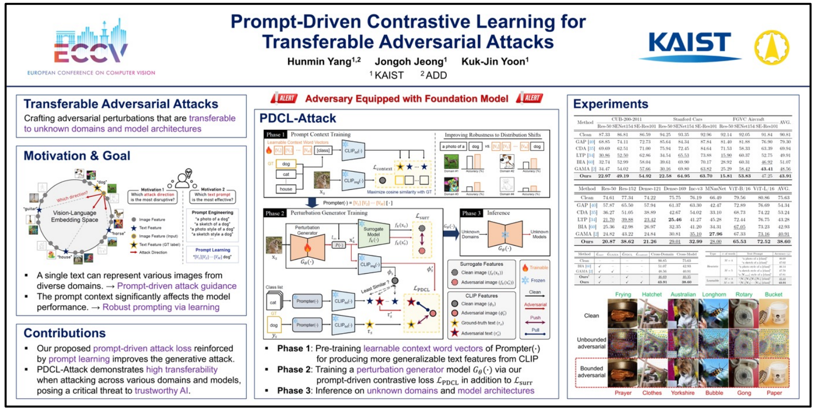
Experimental Results

- Qualitative results of image classification



Conclusion

- PDCL-Attack demonstrates high attack transferability across unknown domains and model architectures, posing a critical threat to trustworthy AI.
- CLIP model guidance reinforced by prompt learning in a joint vision-language space significantly enhances the attack transferability.
- We hope our work inspires further research on training robust models to defend against adversaries equipped with emerging foundation models.



Poster ID : #11 (10:30-12:30)

Project Page