ICCV OCT 19-23, 2025 HONOLULU HAWAII

网络空间安全学院 School of Cyberspace Science and Technology

Boosting Generative Adversarial Transferability with Self-supervised Vision Transformer Features

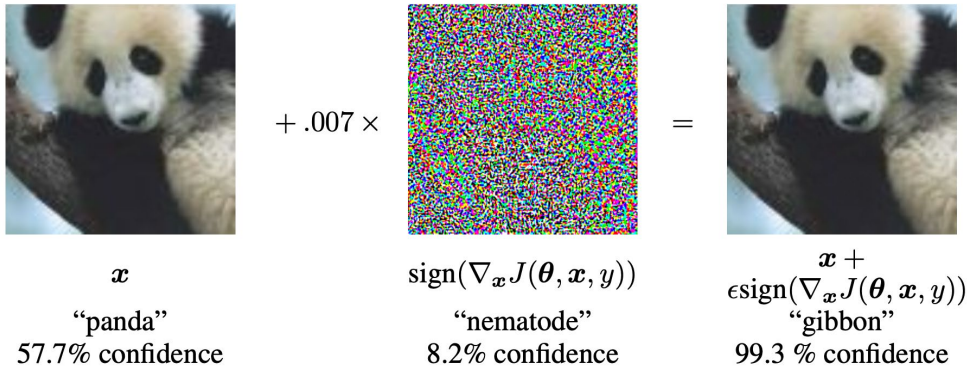
Shangbo Wu[1], Yu-an Tan[1], Ruinan Ma[1], Wencong Ma[2], Dehua Zhu[1], Yuanzhang Li[2*]

[1]*School of Cyberspace Science and Technology* [2]*School of Computer Science and Technology*

Beijing Institute of Technology

What are adversarial attacks and transferability?

- **Adversarial examples** are inputs with subtle modifications that intentionally fool deep neural networks (DNNs) into incorrect predictions while appearing unchanged to humans.
- The transferability of adversarial examples[1] drive real-world black-box attacks on DNNs without the adversary's access to their internals.



▲ One of the first and most popular adversarial attacks to date: the **Fast Gradient Sign Method (FGSM)**[2].

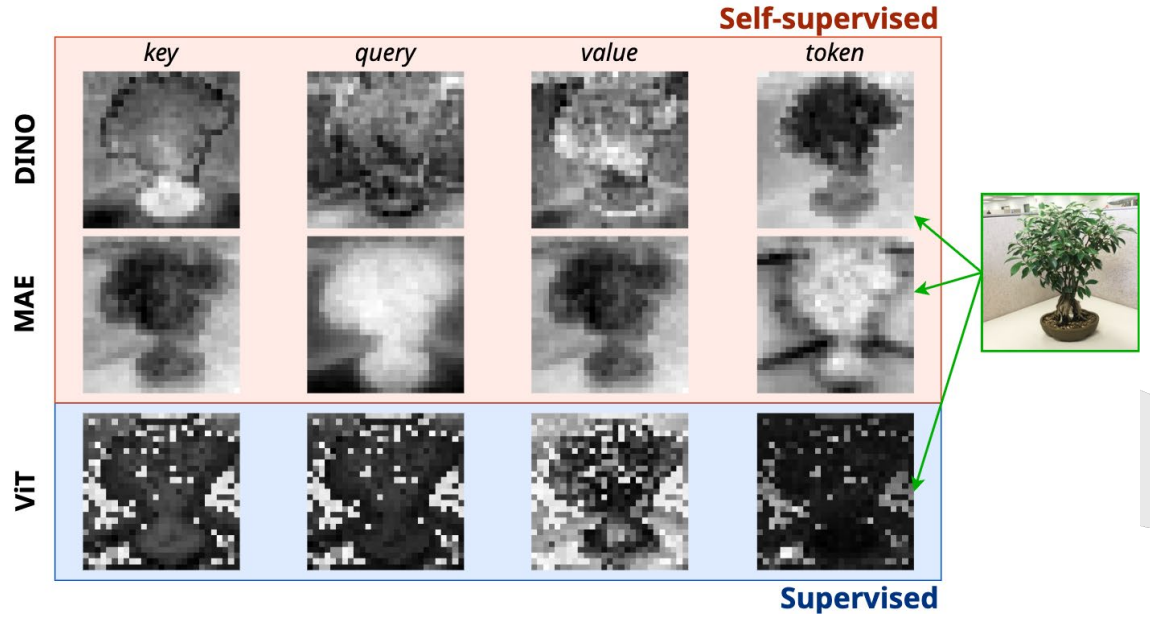
▲ A printable **adversarial patch** brings adversarial attacks into the real-world[3].

[1] Liu, et al. "Delving into Transferable Adversarial Examples and Black-box Attacks." *International Conference on Learning Representations*. 2016.
[2] Goodfellow, et al. "Explaining and Harnessing Adversarial Examples." International Conference on Learning Representations, 2015.
[3] Brown, et al. "Adversarial Patch." 31st Conference on Neural Information Processing Systems, 2017.

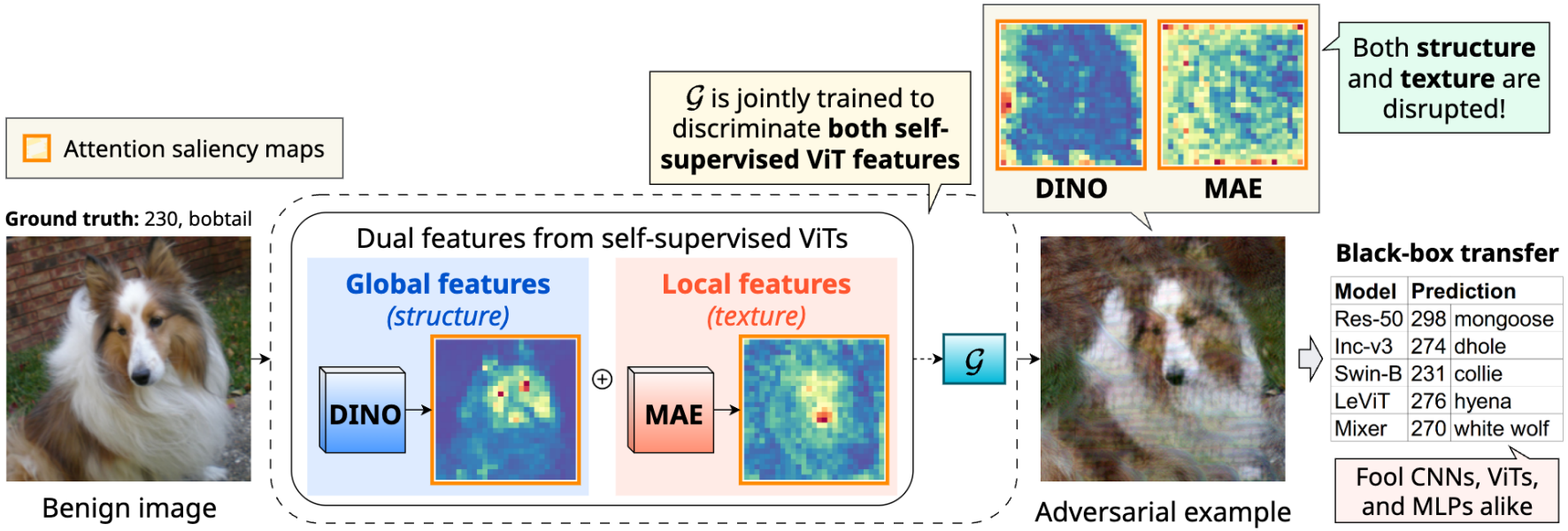
The missing piece in feature-space attacks

- Feature-space attacks improve transferability by disrupting shared model internal features.
  - **Gap 1 -** Existing methods target features from supervised ConvNets and rely on label-wise loss.
  - **Gap 2 -** Features are extracted from whole intermediate layers.
- **Self-supervision** synergies well with the **Transformer** architecture!
  - Offers cleaner, more generalizable representations.
- This suggests more surgical targets within ViTs, especially trained with self-supervision.

► ViT features learnt by **self-supervision** (DINO and MAE) are **less noisy, more expressive** than those by **supervision**.
► Internal feature facets **(key/query/value)** reveal less noisy, more distinct details than token-level outputs.



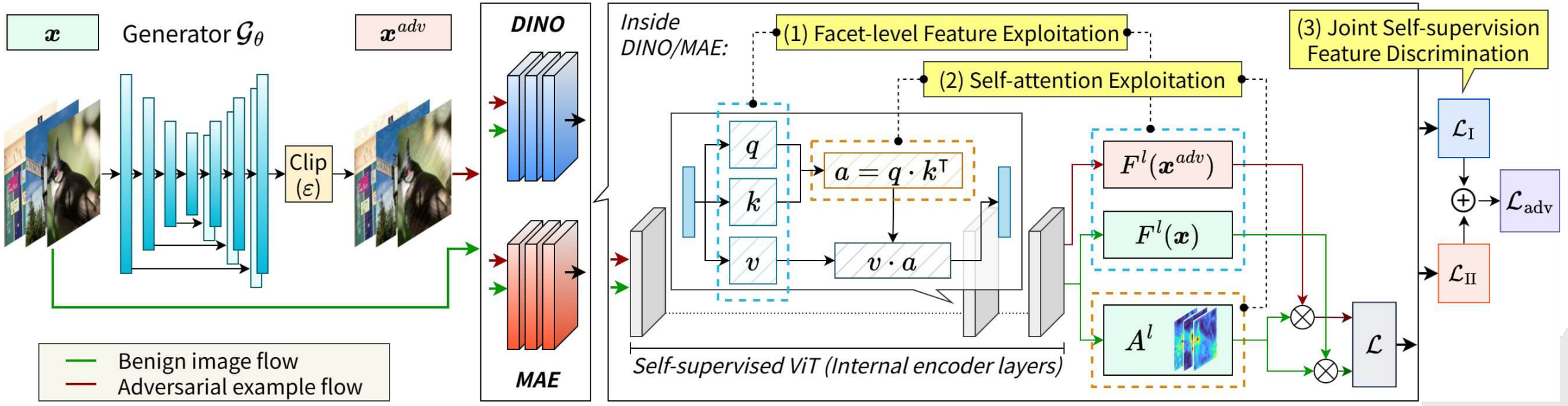
How to utilize or exploit self-supervised ViT features?



- Our idea – **dSVA** (generative dual self-supervised ViT features attack)
  - **Target facet-level features:** we extract representations from $q/k/v$ feature facets to exploit.
  - **Exploit self-attention:** we apply self-attention saliency as feature landmark guidance.
  - **Jointly disrupt self-supervised features:** we consider the two branches of self-supervision paradigms – **contrastive learning, CL** (DINO) and **masked image modelling, MIM** (MAE).

dSVA – framework overview

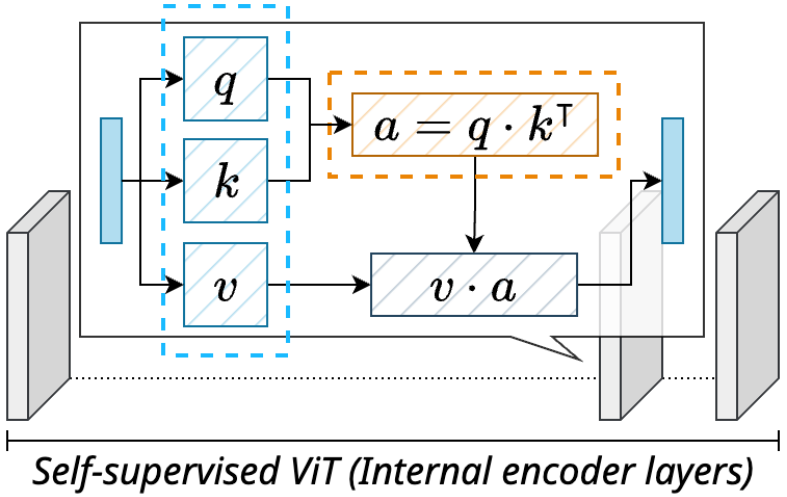
- **Goal:** generator G_θ learns to create $x_{adv} = G_\theta(x)$ under perturbation budget ε, s.t. deep features of self-supervised ViTs (DINO and MAE) are maximally disrupted in a facet-aware, attention-guided way.



dSVA – facet-level feature exploitation

- Instead of attacking whole intermediate layer tokens, dSVA disrupts internal facets of the self-supervised ViT's multi-head self-attention, i.e., queries, keys, and values.

► Inside the Transformer encoder block:
LayerNorm – MSA – MLP
► Inside MSA: **linear projections to q, k, v** from previous layer tokens, and the fused **t**.



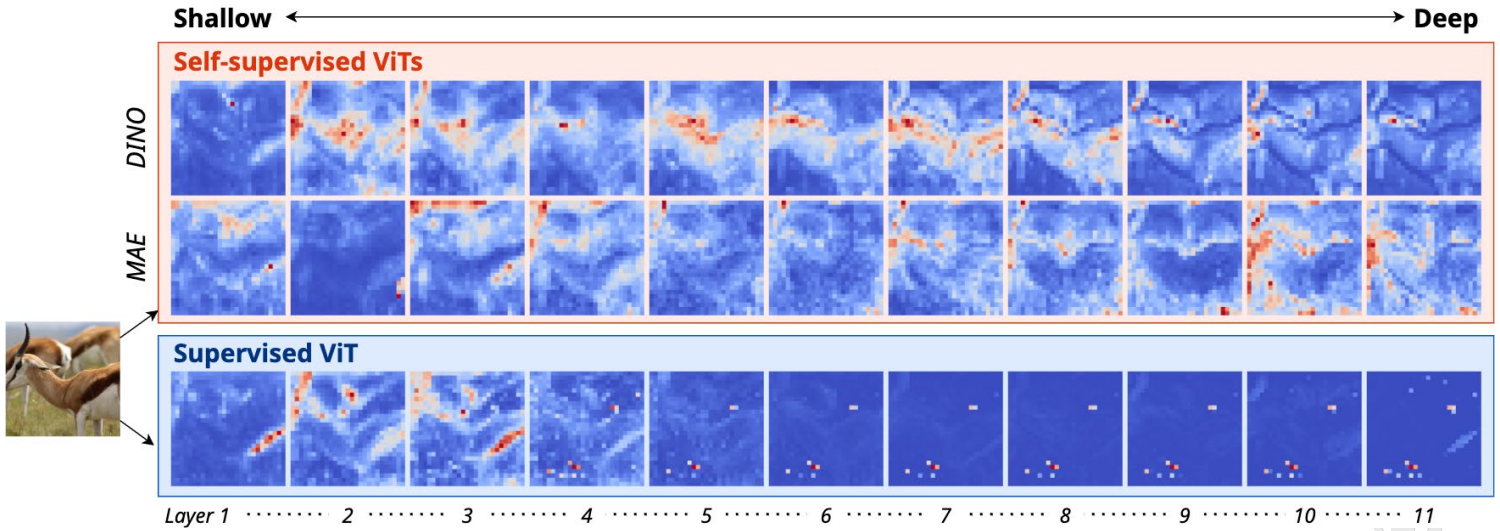
(As with all generative adversarial attacks…)

- Goal is to minimize cosine similarity between benign and adversarial facet features at a chose layer l, so as to make deep representations diverge.

dSVA – self-attention exploitation

- Self-supervised ViTs produce clean and semantically meaningful attention saliency maps.
- Turn them into dense guidance to focus the generator on disrupting impactful features.

► Attention visualized in **self-supervised ViTs** (DINO and MAE) are **less noisy** and **capture various levels of semantics**.
► Representations in supervised ViTs collapse into **homogeneous primitive patterns**.



- Attention map at layer l: select from the $[cls]$ token to all patch tokens across all heads, and average over heads to get a 2D token map S^l.

$$A^l_{[\mathrm{CLS}]} = A^l[:, :, 0, 1:].$$

$$S^l = \frac{1}{H}\sum_{h=1}^{H} A^l_{[\mathrm{CLS}]}[h],$$

dSVA – joint self-supervision feature discrimination

- Two paradigms for self-supervision: contrastive learning, and masked image modelling.
  - Contrastive learning (CL, with DINO) captures global, long-range structural cues.
  - Masked image modeling (MIM, with MAE) focuses on local textural detail.
- Jointly training to exploit both structure and texture yields the most transferable perturbations across architectures and defenses!

$$\mathcal{L}_{\text{adv}} = \lambda \cdot \mathcal{L}_{\text{I}} + (1 - \lambda) \cdot \mathcal{L}_{\text{II}},$$

DINO's L_I MAE's L_{II}

$$\mathcal{L} = \arg\min_{\boldsymbol{\theta}} \mathcal{D}_{\cos} \left(F^l \left(\boldsymbol{x} \right) \odot \left(\gamma \cdot S^l \right), F^l \left(\boldsymbol{x}^{adv} \right) \odot \left(\gamma \cdot S^l \right) \right).$$

Evaluation

- **Training dataset:** training set of ImageNet (1.28 mil) used for training.
- **Test dataset:** NeurIPS 2017 Adversarial Learning dataset (1000 images).
- **Implementation details:**
  - ViT-B/16 architecture for both DINO and MAE.
  - ResNet generator.
  - Perturbation budget $\varepsilon = 10$.
- **Competitors:**
  - Generative attacks: **CDA, BIA** (w/ surrogate VGG-19, ResNet-152, DenseNet-169, and **ViT-B/16**).
  - Feature-level gradient-based attacks: **PNA, TGR, ATT**.
  - Classic attack: **MI-FGSM**.

Results – transferability across black-box models

- dSVA with joint self-supervised features, without using labels, dominates across supervised ConvNets, ViTs, and MLPs.
- dSVA maintains competitive performance on structurally matched ConvNet baselines.
- Findings: existing feature-wise attacks cannot take full advantage of ViTs without dSVA's exploitation schemes.

Structurally matched ConvNets

Attack	VGG-16	Res-50	Den-121	Eff-B0	Inc-v3	Inc-v4	Swin-B	MaxViT	PiT-B	Visformer	LeViT	Mixer
CDA (VGG-19)	99.31	69.23	59.19	76.38	52.94	61.96	16.53	14.63	9.48	32.40	29.79	23.02
CDA (Res-152)	92.98	88.88	87.02	75.32	63.85	74.97	11.82	7.78	5.86	39.03	35.85	22.78
CDA (Den-169)	92.98	87.63	97.03	90.96	67.59	78.94	26.88	22.41	20.98	69.67	65.11	52.01
BIA (VGG-19)	97.58	74.32	84.93	77.77	66.63	76.96	19.35	15.25	12.46	34.68	35.96	27.53
BIA (Res-152)	94.94	92.52	86.47	65.11	62.46	81.37	22.18	17.32	11.40	45.55	29.15	29.60
BIA (Den-169)	93.67	86.07	95.49	81.17	75.40	71.78	17.36	9.44	10.65	32.71	44.47	38.98
CDA (ViT-B/16)	92.75	74.32	90.10	87.23	81.82	82.25	62.13	33.09	59.74	78.05	85.20	80.63
BIA (ViT-B/16)	52.93	21.83	33.77	32.13	31.55	34.62	8.89	5.50	6.39	17.81	27.34	40.68
MI (ViT-B/16)	52.59	32.33	47.85	52.34	38.07	35.61	49.69	31.02	42.92	47.31	43.51	65.16
PNA (ViT-B/16)	46.49	33.99	42.68	50.64	37.97	36.05	50.84	35.68	46.96	51.04	51.49	74.30
TGR (ViT-B/16)	54.89	35.14	51.60	57.02	37.54	40.35	51.15	34.02	45.26	50.72	46.38	79.78
ATT (ViT-B/16)	60.41	40.85	56.55	64.47	43.32	44.43	59.10	40.15	51.12	58.80	56.02	82.52
dSVA (DINO)	86.54	57.59	83.17	88.51	78.50	78.61	33.05	21.27	35.04	72.67	67.41	78.81
dSVA (MAE)	94.36	78.07	86.36	84.04	77.75	79.71	47.38	31.85	33.55	63.25	64.32	56.64
dSVA (Joint)	96.78	81.70	94.83	95.32	89.73	91.73	59.83	41.29	50.48	81.37	85.21	85.38

▲ dSVA's joint variant achieves state-of-the-art transferability – **+13.70% avg** over dSVA's single-model variants.

Results – transferability to defense models

- dSVA offers superior performance across all robust models, even with SOTA defenses.
- Even robust models require essential features, dSVA destroys them at a generalized level.

Attack	Inc-v3$_{adv}$	Inc-v3$_{ens3}$	Inc-v4$_{ens4}$	IncRes-v2$_{ens}$	IncRes-v2$_{adv}$	Eff-b0$_{ap}$
CDA (VGG-19)	25.05	16.36	9.78	10.73	34.90	67.39
CDA (Res-152)	43.01	38.60	28.88	29.27	61.89	73.91
CDA (Den-169)	53.44	41.11	27.08	24.58	66.00	83.33
BIA (VGG-19)	39.57	28.35	21.24	17.60	62.19	79.71
BIA (Res-152)	32.26	27.15	19.89	17.50	63.29	70.29
BIA (Den-169)	55.91	43.40	37.64	30.52	59.08	86.23
CDA (ViT-B/16)	65.91	53.98	50.67	38.54	71.11	86.23
BIA (ViT-B/16)	22.80	15.38	12.02	10.83	24.97	52.17
MI (ViT-B/16)	26.67	22.46	21.91	18.85	26.98	55.07
PNA (ViT-B/16)	27.63	22.90	22.70	19.79	29.69	55.07
TGR (ViT-B/16)	30.22	25.85	24.83	21.67	29.89	67.39
ATT (ViT-B/16)	40.43	36.21	33.03	29.79	41.52	75.36
dSVA (DINO)	66.13	54.09	49.33	43.85	75.03	**89.96**
dSVA (MAE)	50.11	32.39	28.88	23.85	66.70	76.09
dSVA (Joint)	**79.03**	**68.16**	**62.70**	**52.50**	**88.06**	89.13

Attack	Res-18 [48]	Res-50 [63]	ViT-B [39]	Swin-B [39]	XCiT-S12 [13]	ViT-S +ConvStem [51]	ConvNeXt +ConvStem [51]	ConvNeXt-v2+Swin-L [3]
CDA (VGG-19)	7.13	8.25	6.09	10.15	7.91	6.69	4.96	5.68
CDA (Res-152)	12.56	11.39	12.31	13.20	10.74	7.39	7.04	7.07
CDA (Den-169)	11.21	12.54	9.96	16.38	13.93	10.33	8.19	8.89
BIA (VGG-19)	12.05	11.22	8.85	12.96	11.22	9.51	7.50	7.50
BIA (Res-152)	16.13	15.35	14.52	19.32	16.06	11.97	10.61	8.24
BIA (Den-169)	14.09	14.19	18.95	22.62	16.65	10.92	9.80	9.42
CDA (ViT-B/16)	12.39	13.04	8.85	18.70	14.52	11.39	9.00	8.67
BIA (ViT-B/16)	10.70	9.90	12.86	12.47	8.97	8.10	7.50	5.03
MI (ViT-B/16)	7.81	7.92	11.62	12.96	8.26	7.51	6.46	6.96
PNA (ViT-B/16)	7.13	8.58	10.79	14.06	8.03	7.98	6.11	7.71
TGR (ViT-B/16)	12.73	11.55	16.18	18.34	12.16	11.50	8.88	9.32
ATT (ViT-B/16)	12.22	12.05	17.70	19.19	12.04	11.27	8.65	10.49
dSVA (DINO)	**20.88**	19.47	**23.93**	**26.28**	21.49	**15.96**	**12.80**	11.67
dSVA (MAE)	15.11	14.69	14.52	18.46	15.94	11.50	10.04	10.39
dSVA (Joint)	19.19	**19.64**	21.44	24.45	**22.31**	14.79	12.11	**11.99**

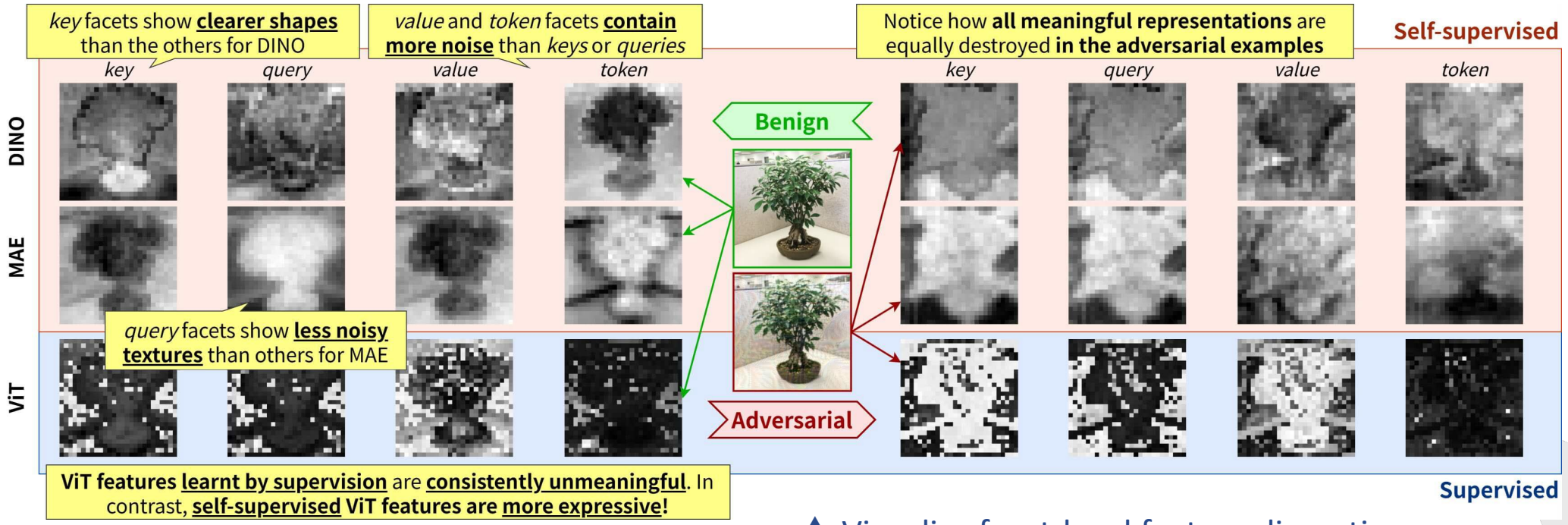
▲ dSVA's joint variant exceeds baselines by **+32.98% avg.**

▲ dSVA excels at evading **even SOTA defenses** (These models come from **robustbench**[1]).

[1] Croce, Francesco, et al. "RobustBench: a standardized adversarial robustness benchmark." *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.*

Results – qualitative visualizations

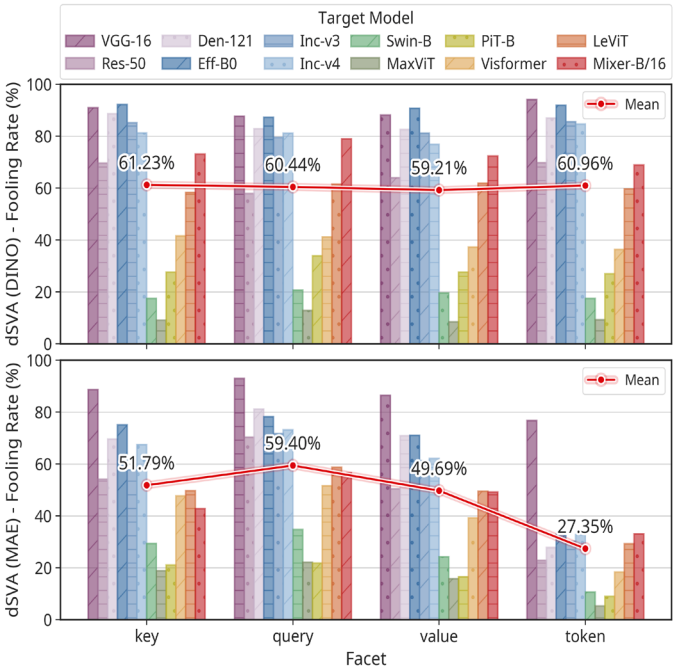
- PCA projections of facet-level features before v.s. after the attack.
- Self-supervised ViT features are richer and less noisy than supervised.
- dSVA disrupts both structural + textural semantics consistently.



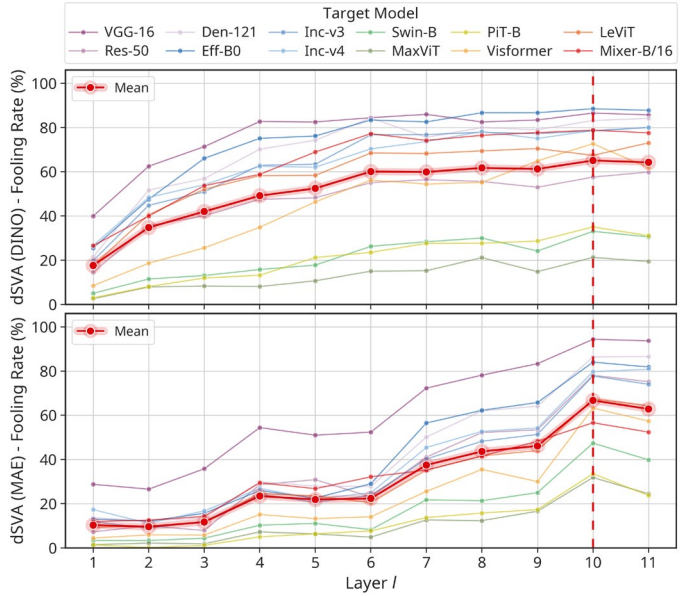
▲ Visualize facet-level feature disruptions, comparison between self-supervised v.s. supervised.

Results – analysis on parameter impact

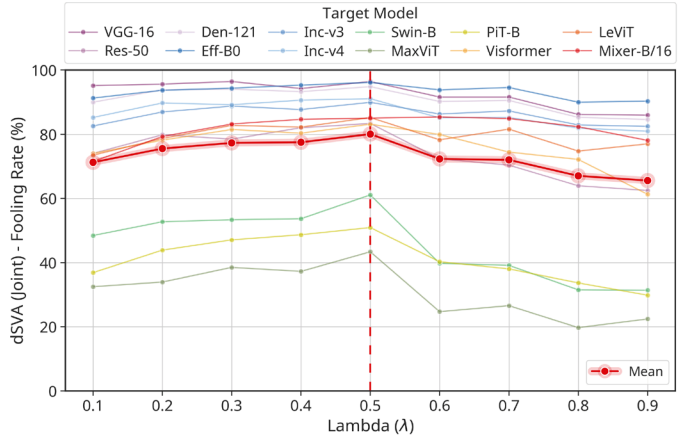
- Feature facet: q or k facet more robust than v or t (token).
- Layer: penultimate best ($l = 10$), last layer ($l = 11$) drops generalizability.
- Joint training parameter: sweet point lies in $\lambda = 0.5$, i.e., middle point.



▲ Facet choice (q, k, v, or t)



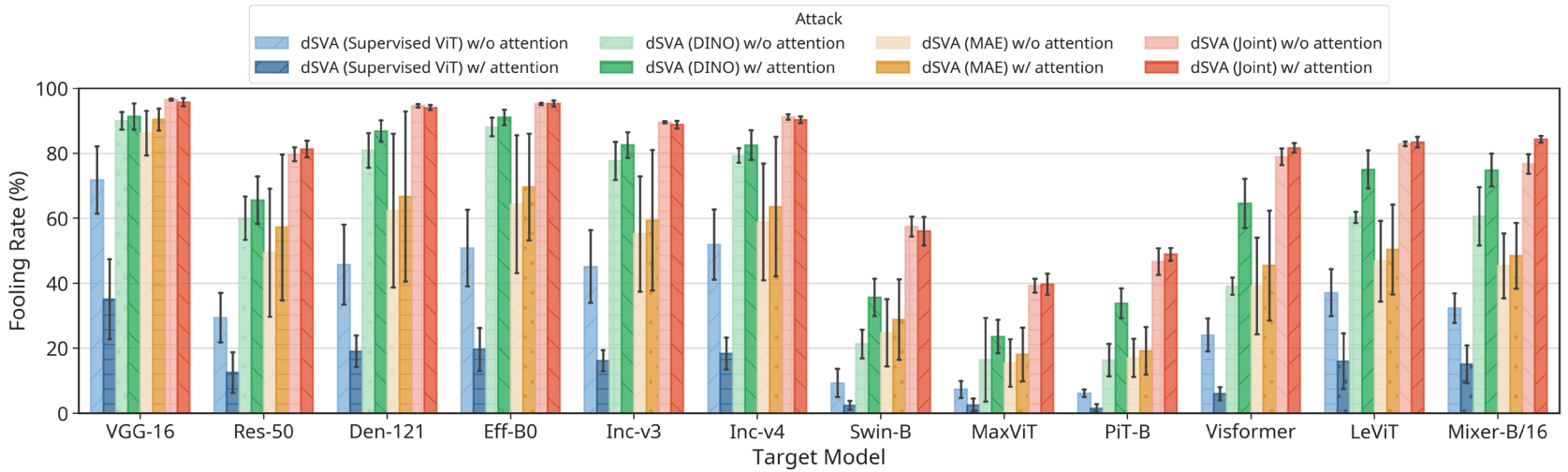
▲ Layer choice ($l = 1$ to 11)



▲ λ choice (from 0.1 to 0.9)

Results – ablations

- Compare dSVA …
  - With supervised ViT-B/16 as surrogate v.s. using DINO/MAE.
  - Applying attention saliency regularization or not.
- dSVA always performs better with self-supervised ViT features.
- Supervised ViT's attention impairs adversarial effectiveness.
- dSVA (Joint) works best with attention regularization active when targeting more sophisticated models.



Results – cross-domain transferability

- dSVA delivers **+6% avg improvements** in most cross-domain cases v.s. strong generative attack competitors (CDA/BIA).

Attack	s	A	Domain			
			CIFAR-10	CIFAR-100	SVHN	STL-10
CDA (VGG-19)	/	/	12.65	30.79	3.36	7.56
CDA (Res-152)	/	/	10.34	28.23	5.49	6.15
CDA (Den-169)	/	/	27.42	53.22	6.84	10.31
BIA (VGG-19)	/	/	**39.04**	**68.25**	6.38	9.84
BIA (Res-152)	/	/	26.24	49.36	3.75	7.35
BIA (Den-169)	/	/	22.05	45.82	**12.79**	**10.75**
dSVA (DINO)	16	w/o	13.98	37.67	**12.88**	11.07
dSVA (DINO)	8	w/o	24.05	53.00	6.54	11.18
dSVA (DINO)	16	w/	13.34	37.42	9.30	**12.66**
dSVA (DINO)	8	w/	21.94	48.94	7.53	10.70
dSVA (MAE)	16	w/o	16.89	35.80	6.80	10.41
dSVA (MAE)	8	w/o	24.77	41.15	9.13	10.26
dSVA (MAE)	16	w/	17.47	34.32	4.91	9.31
dSVA (MAE)	8	w/	24.30	44.61	6.74	11.44
dSVA (Joint)	16	w/o	23.64	50.28	8.94	11.04
dSVA (Joint)	8	w/o	**26.87**	**55.53**	8.83	12.42
dSVA (Joint)	16	w/	21.56	43.25	8.82	11.89
dSVA (Joint)	8	w/	24.13	46.73	11.73	11.95

▲ Cross-domain transferability (towards coarse-grained classification domains)

Attack	s	A	CUB-200-2011			Stanford Cars			FGVC Aircraft		
			Res-50	SENet154	SE-Res-101	Res-50	SENet154	SE-Res-101	Res-50	SENet154	SE-Res-101
CDA (VGG-19)	/	/	29.49	29.94	20.79	21.84	20.95	10.42	24.81	40.91	23.02
CDA (Res-152)	/	/	49.85	48.77	34.77	48.08	37.91	21.60	33.80	48.01	36.19
CDA (Den-169)	/	/	39.55	29.52	36.40	42.16	25.26	19.22	30.61	32.92	33.77
BIA (VGG-19)	/	/	62.21	52.78	36.84	70.93	37.01	29.86	82.61	51.17	51.27
BIA (Res-152)	/	/	63.53	68.15	38.92	56.91	58.49	19.03	41.52	77.61	42.33
BIA (Den-169)	/	/	**83.36**	65.75	45.77	**91.67**	51.75	52.57	**96.16**	59.78	65.22
dSVA (DINO)	16	w/o	38.86	51.65	43.66	**53.57**	59.22	50.79	**72.52**	81.45	64.73
dSVA (DINO)	8	w/o	71.18	61.15	59.57	49.39	59.76	**56.23**	54.38	77.71	67.96
dSVA (DINO)	16	w/	41.55	49.48	47.75	47.01	51.25	47.23	53.57	61.83	66.10
dSVA (DINO)	8	w/	33.68	40.99	38.12	33.78	37.92	29.92	37.12	46.25	55.68
dSVA (MAE)	16	w/o	42.93	51.81	37.56	28.80	47.10	20.24	34.13	50.62	43.86
dSVA (MAE)	8	w/o	37.38	58.97	36.44	44.28	38.30	26.74	29.70	50.10	36.58
dSVA (MAE)	16	w/	60.08	63.80	42.42	41.22	62.48	26.79	38.81	72.95	57.45
dSVA (MAE)	8	w/	42.38	62.11	41.99	46.04	38.99	29.33	30.41	52.90	43.73
dSVA (Joint)	16	w/o	**78.77**	79.62	66.11	48.67	**68.47**	51.97	65.65	89.24	**83.15**
dSVA (Joint)	8	w/o	62.58	72.17	59.11	41.42	55.68	41.17	46.76	75.07	63.62
dSVA (Joint)	16	w/	76.44	**79.64**	**69.72**	47.29	67.91	50.99	68.94	**89.93**	77.37
dSVA (Joint)	8	w/	70.88	78.85	68.24	47.25	66.30	50.12	68.15	87.97	74.10

▲ Cross-domain transferability (towards fine-grained classification domains.

Conclusion

dSVA – a generative adversarial attack that successfully exploits deep features distilled through the self-supervised learning of ViTs.

Key takeaways:

- Dual exploitation of visual structure (**CL**, from DINO) and texture features (**MIM**, from MAE) – brings complementary improvements.
- Facet-level self-supervised ViT features (q/k facets) – powerful fine-grained attack targets better than whole-layer tokens, no class or labels required.
- Self-attention saliency guidance – further pushes attack effectiveness, especially on newer architecture DNNs, yielding SOTA transferability and defense evasion.

Thank you!

Shangbo Wu
shangbo.wu@bit.edu.cn