



重慶理工大學
CHONGQING UNIVERSITY OF TECHNOLOGY



Boosting the Transferability of Adversarial Attack on Vision Transformer with Adaptive Token Tuning

Di Ming, Peng Ren, Yunlong Wang, Xin Feng*

School of Computer Science and Engineering, Chongqing University of Technology
Chongqing, China

diming@cqut.edu.cn, misterr_2019@163.com, ylwang@cqut.edu.cn, xfeng@cqut.edu.cn



GitHub Project



AdvML-Group



➤ Background

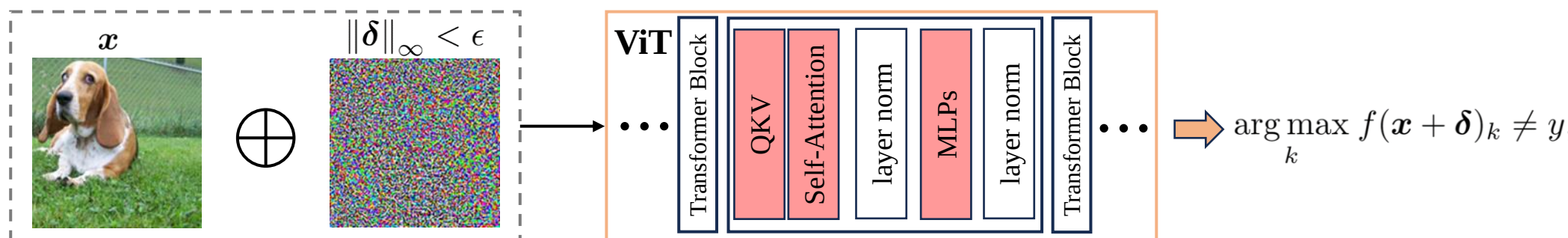
❖ Adversarial Example

- ❑ Crafted by adding tiny perturbations deliberately to benign sample.
- ❑ Aim at disturbing the prediction of deep neural network models, e.g., image classification.

❖ Adversarial Attack on Vision Transformers (ViTs)

- ❑ ViTs demonstrate excellent performance in a range of computer vision tasks.
- ❑ Similar to CNNs, ViTs remain vulnerable to adversarial attacks, which can be described as follows:

$$\arg \max_{\delta} \mathcal{L}(f(x + \delta), y), \quad \text{s.t. } \|\delta\|_{\infty} \leq \epsilon$$



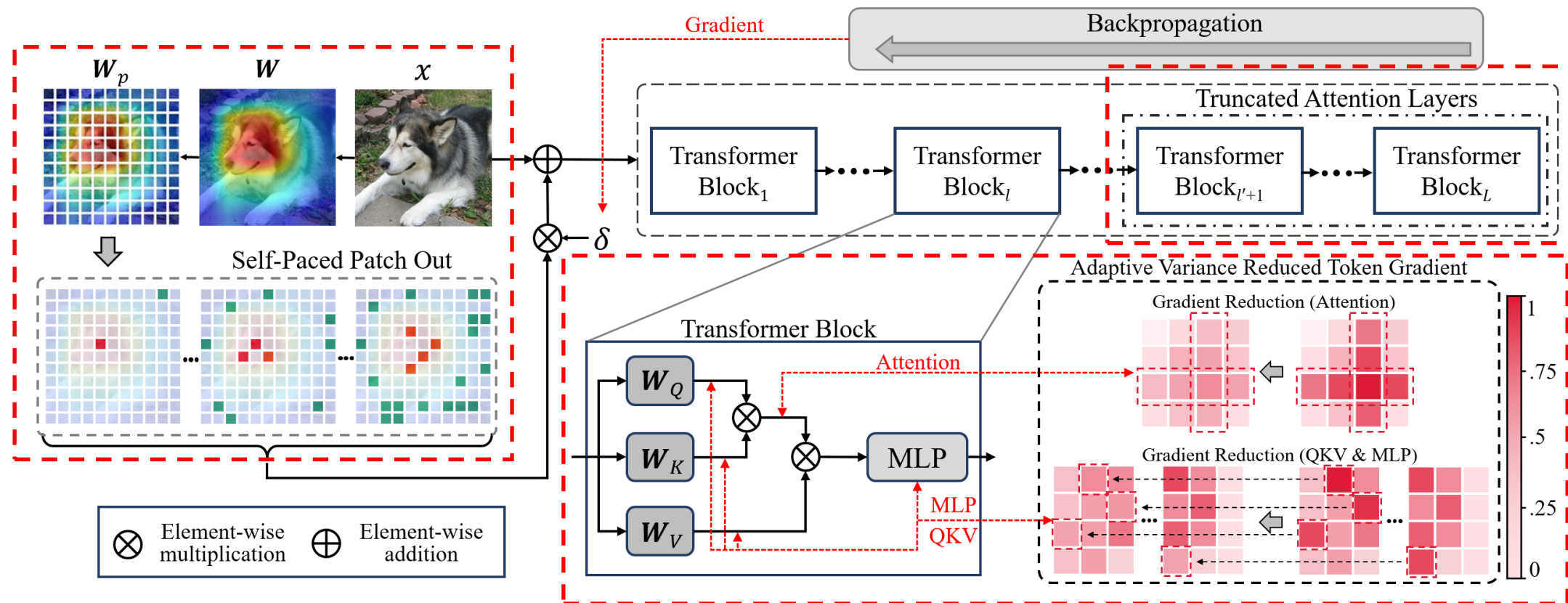
❖ Boosting the Transferability of Adversarial Attack on Vision Transformer with Adaptive Token Tuning (dubbed as ATT-Attack)

- ❑ **Goal:** address the limitations of existing works, e.g., the overly aggressive regularization of token gradient
- ❑ **ATT-Attack:** achieve more transferable and efficient attacks across various target models in black-box setting.



➤ Adaptive Token Tuning (ATT) Attack

❖ Three Optimization Strategies for Improving both Transferability and Efficiency of ViT Attacks.



- ❑ **Adaptive Gradient Re-scaling Strategy:** reduce the overall variance of token gradients.
- ❑ **Self-paced Patch Out Strategy:** enhance the diversity of input tokens.
- ❑ **Hybrid Token Gradient Truncation Strategy:** weaken the effectiveness of attention mechanism.



➤ Adaptive Variance Reduced Token Gradient

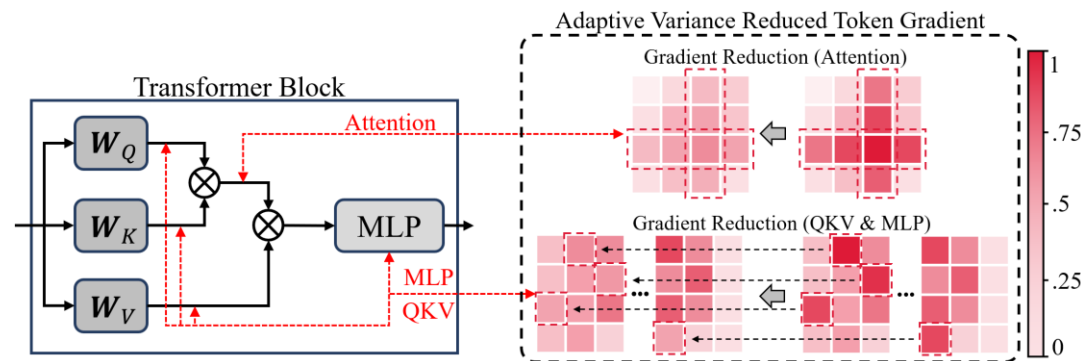
❖ Variance Reduction in a Single ViT Layer

- ❑ The gradient of the i -th token for a given m -module of layer l can be expressed as $\mathbf{g}_i^{(l,m)}$.
- ❑ The maximum gradient is defined as $\arg \max_{i \in \{1, \dots, n\}} \mathbf{g}_i^{(l,m)}$.
- ❑ Mildly re-scale token gradient via gradient penalty factor γ :

$$m = \text{QKV or MLP} \Rightarrow \mathbf{g}_i^{(l,m)} = \gamma \cdot \mathbf{g}_i^{(l,m)}$$

$$m = \text{Attention} \Rightarrow \mathbf{g}_i^{(l,m)} = \gamma \cdot \mathbf{g}_i^{(l,m)}, i \in \mathcal{S}$$

where $\mathcal{S}$ represents the set of extreme token gradients located in the same row or column as the largest token gradient.



❖ Adaptive Variance Reduction Throughout ViT Layers

- ❑ Smooth the variance of token gradients $\Phi_t^{(l,m)} = \text{Var}(\mathbf{g}^{(l,m)})$ between consecutive ViT layers by defining an adaptive gradient updating strategy as:

$$\mathbf{g}_{i,t}^{(l,m)} = \mathbf{g}_{i,t}^{(l,m)} \cdot \left(\gamma + \lambda \left(1 - \sqrt{\Phi_t^{(l,m)} / \Phi_t^{(l+1,m)}} \right) \right)$$

where λ is the adaptive factor balancing the relative importance between the gradient penalty factor and the ratio of gradient variances.



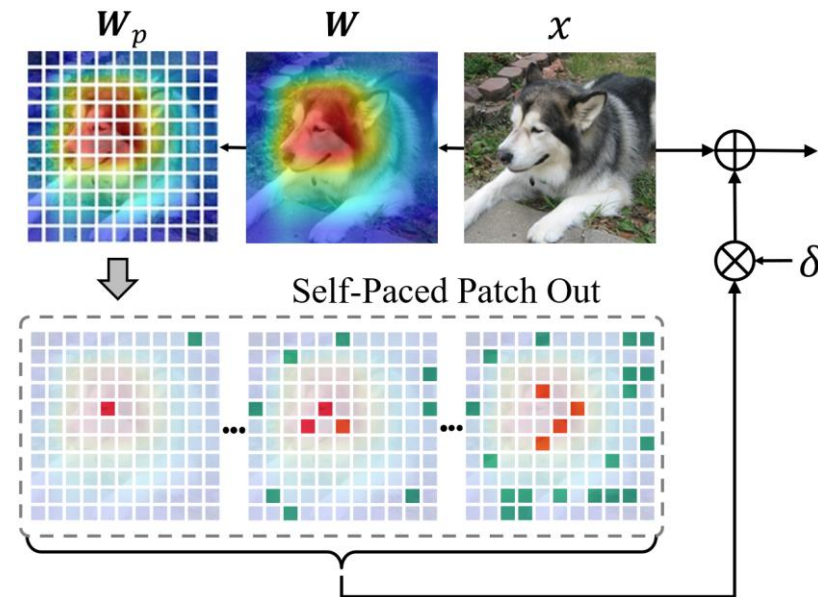
➤ Self-Paced Patch Out under Semantic Guidance

❖ Generating Semantic Guided Sparse Mask

- Based on Grad-CAM, we construct the feature importance matrix from an intermediate ViT layer l as $W = \sum_{i=1}^{C^{(l)}} G_i^{(l)} \odot F_i^{(l)}$.
- According to the partition of the input, we define the patch-level feature importance matrix as $W_p = \{W_p^1, \dots, W_p^n\}$, and measure each patch's importance by the Frobenius norm $\|W_p^i\|_F$.
- By normalizing W_p as c_p^i and introducing α and β to control scaling and offset, the semantic guided sparse mask can be generated by:

$$w = (q < \alpha \cdot c - \beta)$$

where $q \in [0, 1]^{C \times H \times W}$ is a random variable sampled from the patch-level uniform distribution $\mathcal{U}_p(0, 1)$.



❖ Self-Paced Patch Out via Progressive Sparse Mask

- Introducing self-paced patch out strategy to control the number of discarded patches for each iteration at a dynamic pace:

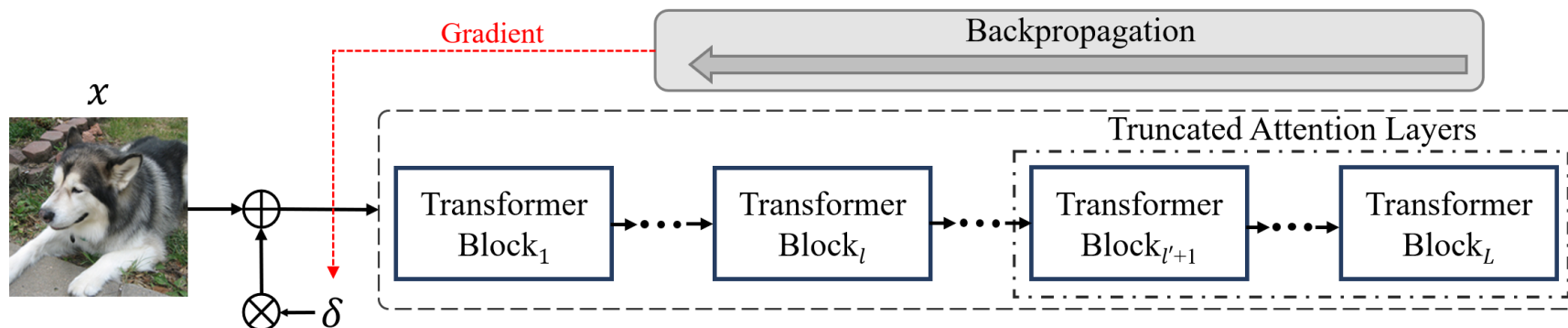
$$w_t = \{w | w = (q_t < \alpha \cdot c_t - \beta), q_t \sim \mathcal{U}_p(0, 1), c_t = 1 - \frac{t}{T}(1 - c)\}$$

$$N_p(w_1) < N_p(w_2) < \dots < N_p(w_T)$$

- As the iteration t increase, c_t becomes to c gradually, and more perturbation patches are discarded to prevent overfitting.



➤ Weakening the Effectiveness of Attention Mechanism



❖ Truncated Attention Layers

- ❑ To mitigate overfitting by excessive global attention, we introduce hard truncation for deep ViT layers, where the token gradient $g_i^{(l,m)}$ is multiplied with a truncation factor τ for module $m = \text{Attention}$:

$$l \in \{1, \dots, l'\} \Rightarrow \tau^{(l,m)} \neq 0$$

$$l \in \{l' + 1, \dots, L\} \Rightarrow \tau^{(l,m)} = 0$$

❖ Hybrid Token Gradient Truncation

- ❑ To effectively balance the influence of different modules on perturbation training, we further introduce the hybrid truncation strategy as:

$$m = \text{Attention} \Rightarrow \mathcal{S}_\tau^{(m)} = \{\tau^{(1,m)}, \dots, \tau^{(l',m)}, 0, \dots, 0\}$$

$$m = \text{QKV or MLP} \Rightarrow \mathcal{S}_\tau^{(m)} = \{\tau^{(1,m)}, \dots, \tau^{(l',m)}, \tau^{(l'+1,m)}, \dots, \tau^{(L,m)}\}$$

- ❑ By setting $\tau^{(l,\text{Attention})} < \max(\tau^{(l,\text{QKV})}, \tau^{(l,\text{MLP})})$ between ViT modules, we can continue to weaken the effectiveness of attention mechanism



➤ Evaluating the Transferability

❖ Threat Models:

ViTs: ☐ ViT-B/16 ☐ PiT-B ☐ CaiT-S/24 ☐ Visformer-S ☐ DeiT-B ☐ TNT-S ☐ LeViT-256 ☐ ConViT-B

CNNs: ☐ Inc-v3 ☐ Inc-v4 ☐ IncRes-v2 ☐ Res-v2

Def-CNNs: ☐ Inc-v3ens3 ☐ Inc-v3ens4 ☐ IncRes-v2adv

❖ Comparative Methods

Gradient-based: ☐ MIM ☐ VMI ☐ SGM ☐ PNA ☐ TGR ☐ ATT (Ours)

Input Diversity-based: ☐ PO ☐ SPPO (Ours)

❖ Comparison with State-of-the-Art Methods

Model	Attack	ViTs	CNNs	Def-CNNs
ViT-B/16	MIM+PO	61.3	31.3	21.7
	VMI+PO	69.1	42.8	30.9
	SGM+PO	64.8	29.2	18.9
	PNA+PO	70.8	42.6	29.9
	TGR+PO	76.0	46.7	33.3
	Ours+PO	77.1	51.7	37.1
	Ours+SPPO	80.3↑	54.1↑	38.7↑

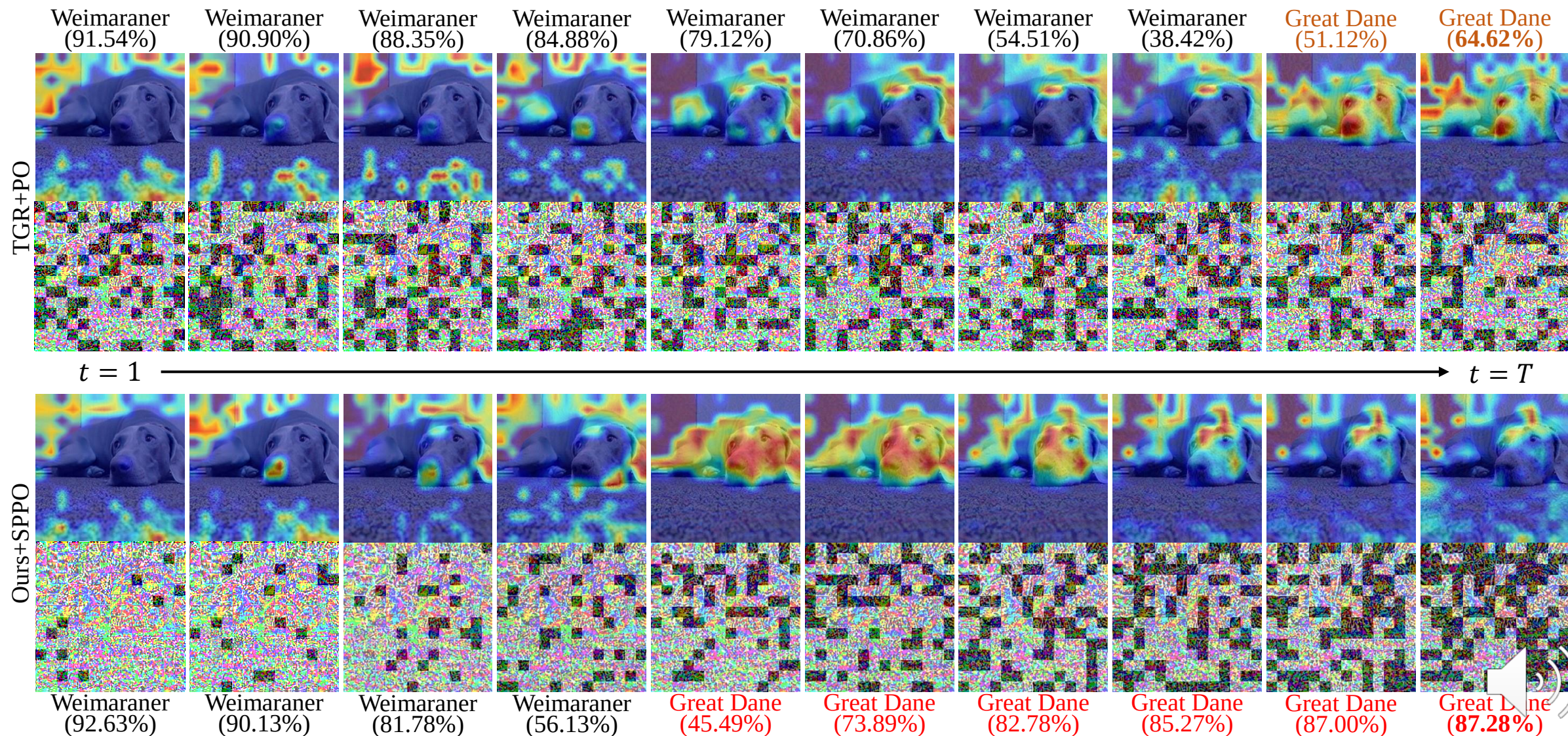
Model	Attack	ViTs	CNNs	Def-CNNs
PiT-B	MIM+PO	47.3	32.5	17.5
	VMI+PO	59.5	46.2	35.8
	SGM+PO	70.0	45.6	21.3
	PNA+PO	73.1	57.8	32.7
	TGR+PO	82.3	68.9	41.3
	Ours+PO	84.2	75.2	48.4
	Ours+SPPO	87.7↑	78.0↑	52.0↑

Model	Attack	ViTs	CNNs	Def-CNNs
CaiT-S/24	MIM+PO	70.3	44.0	29.3
	VMI+PO	76.8	57.8	38.4
	SGM+PO	85.1	49.2	29.3
	PNA+PO	81.6	56.6	39.3
	TGR+PO	88.8	60.5	40.5
	Ours+PO	91.1	71.9	54.3
	Ours+SPPO	92.6↑	75.4↑	58.3↑

Model	Attack	ViTs	CNNs	Def-CNNs
Visformer-S	MIM+PO	54.9	45.7	23.4
	VMI+PO	64.8	56.6	32.6
	SGM+PO	51.6	44.3	15.0
	PNA+PO	68.8	61.8	32.3
	TGR+PO	70.4	64.3	33.5
	Ours+PO	70.5	79.3	44.5
	Ours+SPPO	76.4↑	84.4↑	50.3↑

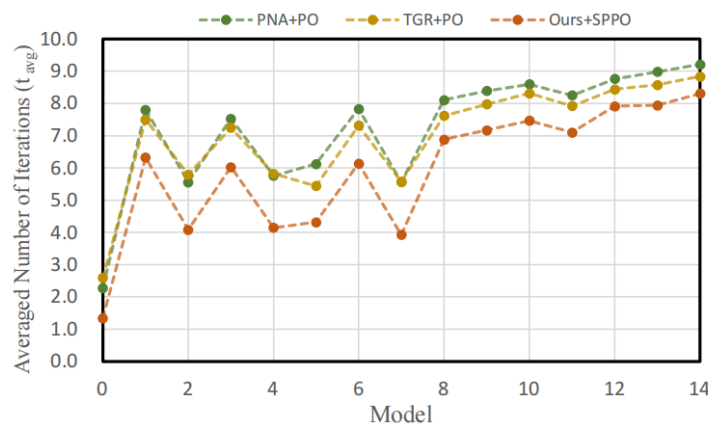


➤ The Analysis of Attack's Effectiveness and Efficiency

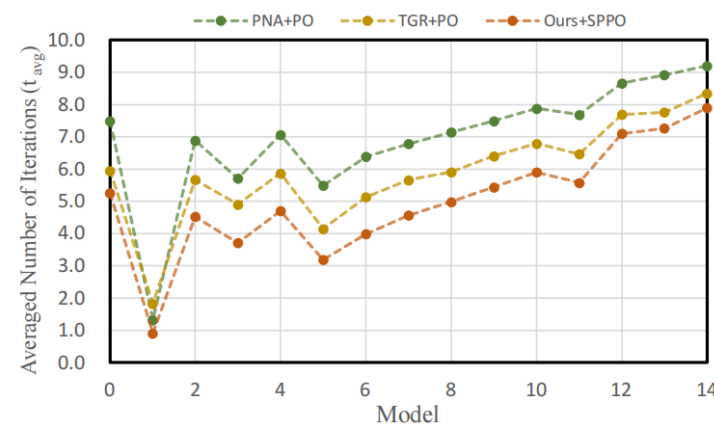


➤ Comparison of Attack Efficiency between Different Methods

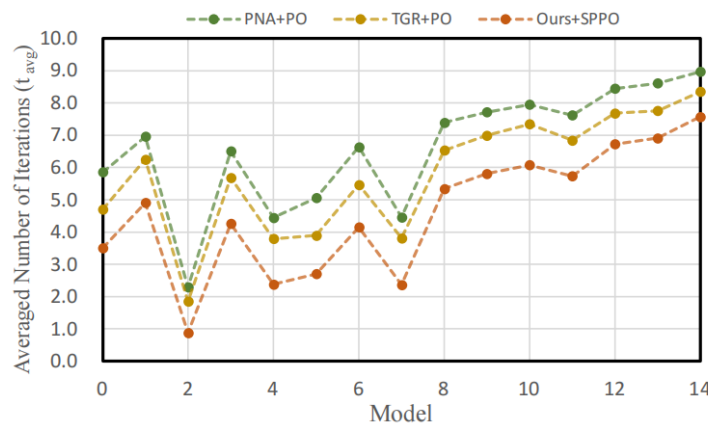
- ❖ We calculate the average of the number of iterations that lead to the first-time misclassification by the model across the entire dataset, defined as $t_{avg} = (1/|D|) \cdot \sum_{i=1}^{|D|} t_i$.



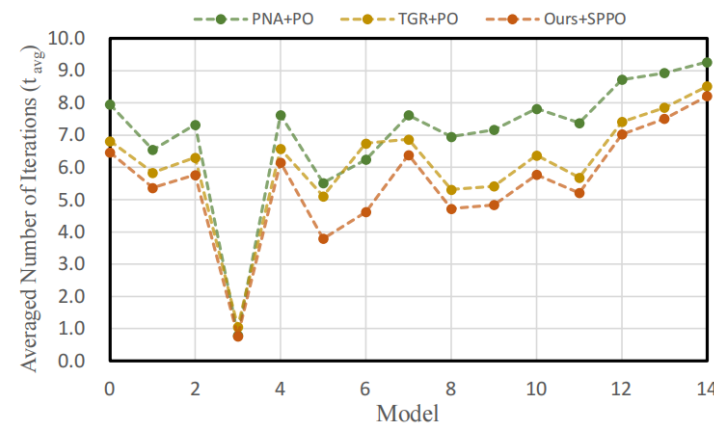
(a) ViT-B/16



(b) PiT-B



(c) CaiT-S-24



(d) Visformer-S



Experiment

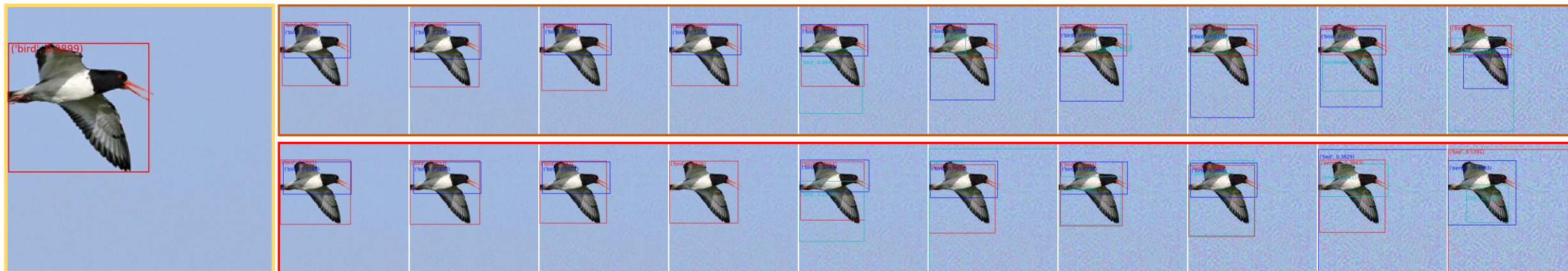


重慶理工大學
CHONGQING UNIVERSITY OF TECHNOLOGY



➤ Evaluating the Transferability of ViT Attack in Cross-Task Scenarios

❖ Transferable ViT Attacks on Object Detection Models



❖ Transferable ViT Attacks on Semantic Segmentation Models

