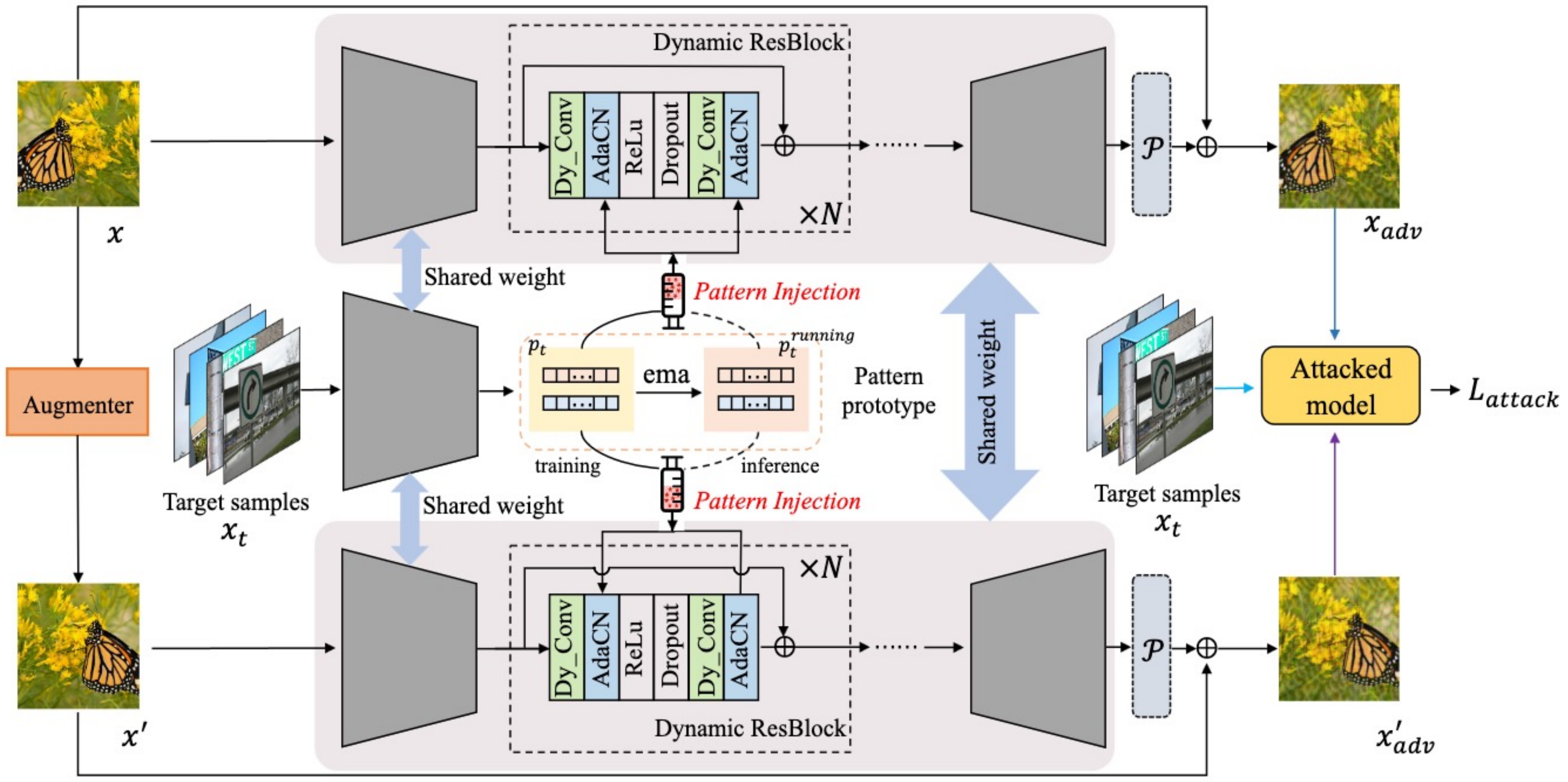
Dynamic Generative Targeted Attacks with Pattern Injection

Weiwei Feng[1], Nanqing Xu[1], Tianzhu Zhang[1,2], Yongdong Zhang[1]
[1]University of Science and Technology of China, [2]Deep Space Exploration Lab
fengww@mail.ustc.edu.cn, xnq@mail.ustc.edu.cn, {tzzhang, zhyd73}@ustc.edu.cn

Poster ID: 386 Tag: WED-PM-386

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Quick Preview



- We propose a dynamic generative model to craft transferable targeted adversarial examples, which can not only inject pattern or style information of the target class to improve transferable targeted attacks, but also learn specialized convolutional kernels for each input instance.
- In the generative model, we design a novel cross-attention guided dynamic convolution module and a pattern injection module.
- We present extensive experiments to demonstrate the effectiveness against normal and robust models.

Introduction

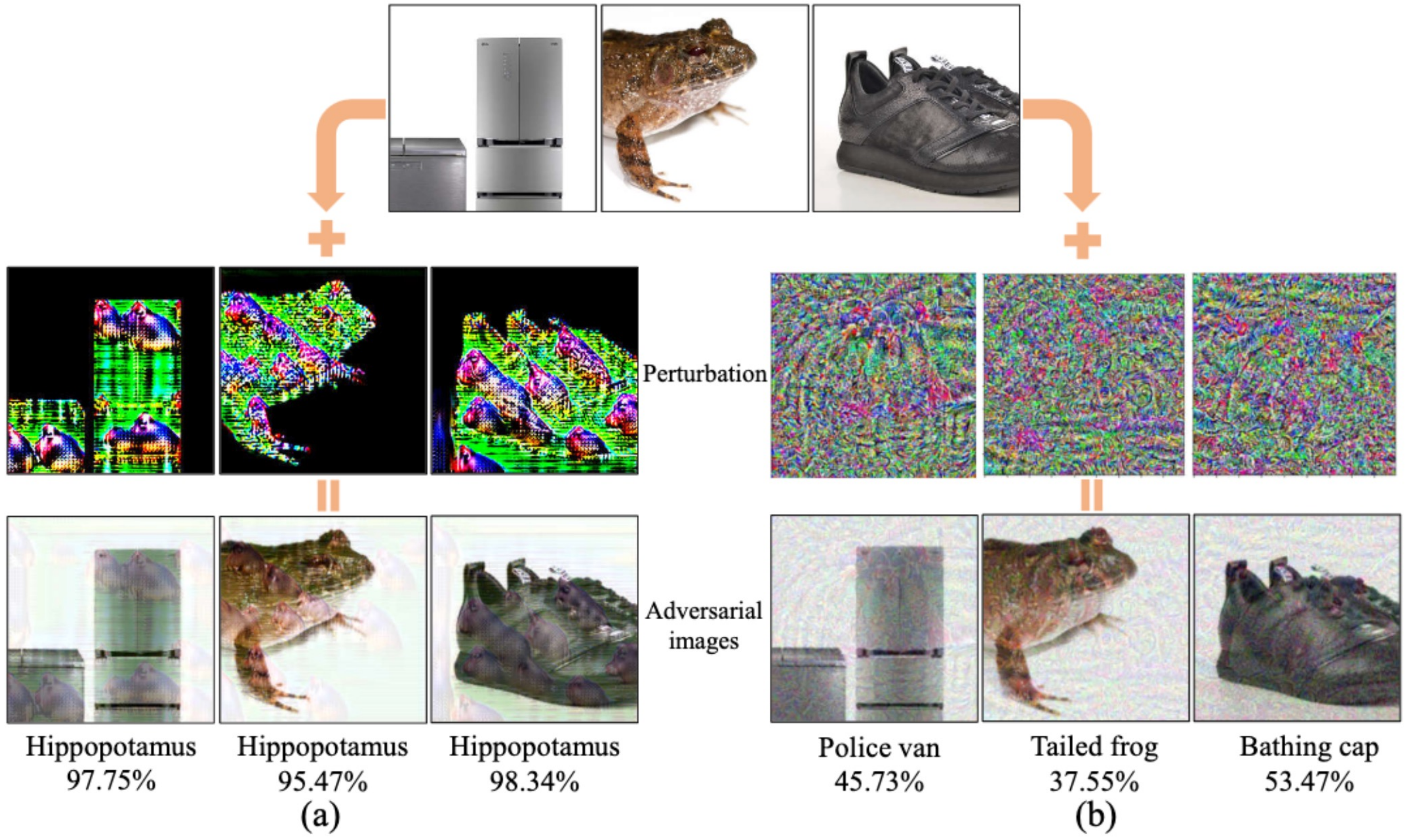
Limitations of current targeted attacks:

- Instance-specific attacks:
  - only take advantage of the specific input instance with iterative gradient updating
  - instance-specific attacks rely on optimizing the classification score of the adversary-desired class label to perturb the specific instance, which ignore the global data distribution
  - lead to adversarial examples over-fitting the white-box model and result in modest transferability of targeted attacks
- Instance-agnostic attack (Generative attacks):
  - Most generative attacks still rely on the target label and the classification boundary information of white-box models rather than the realistic data distribution of the target class.
  - Existing generative attacks apply the same network weights to every input instance in the test dataset.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

Observation:



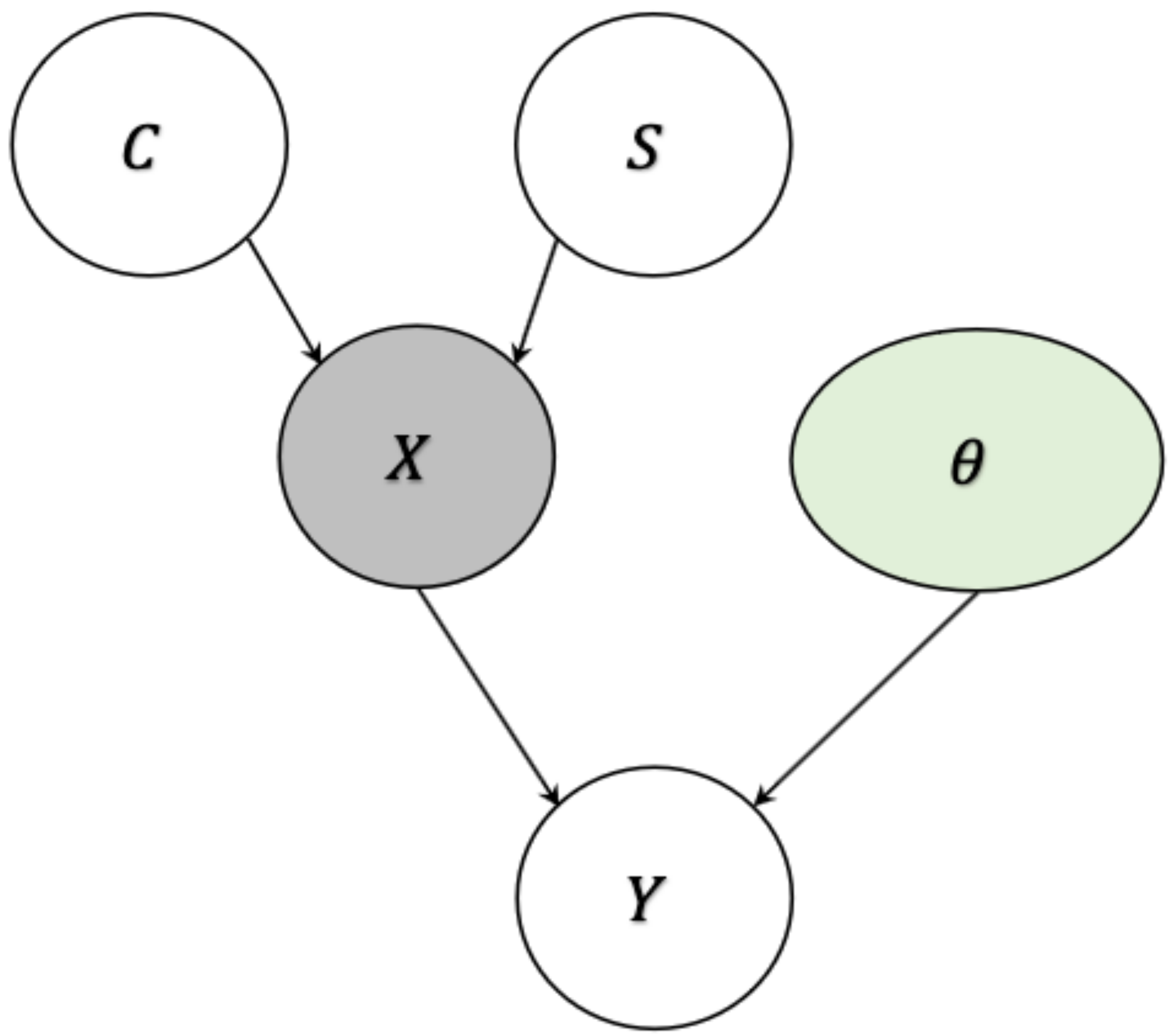
Visualization comparison between adversarial examples generated by our method (a) and the instance-specific method MIM (b). Our perturbations (a) not only show an underlying dependency with the input instance, but also have strong semantic patterns or styles of the target class ("Hippopotamus"). In contrast, the perturbations generated by MIM perform like random noises.

- underlying dependency with the input instance
- strong semantic patterns or styles of the target class

→ More transferable adversarial examples

Motivation

Analysis:



The casual graph of model inference. Each node is a random variable, where C, S, x, y and θ represent content, style or pattern, the input image, the prediction label, and model parameters, respectively.

- For the input images x, we propose to group the whole causes of x into content cause C and style cause S
- We expand the prediction $P_\theta\ (y \mid x)$ and $P_\theta\ (y \mid x_{adv})$ as:

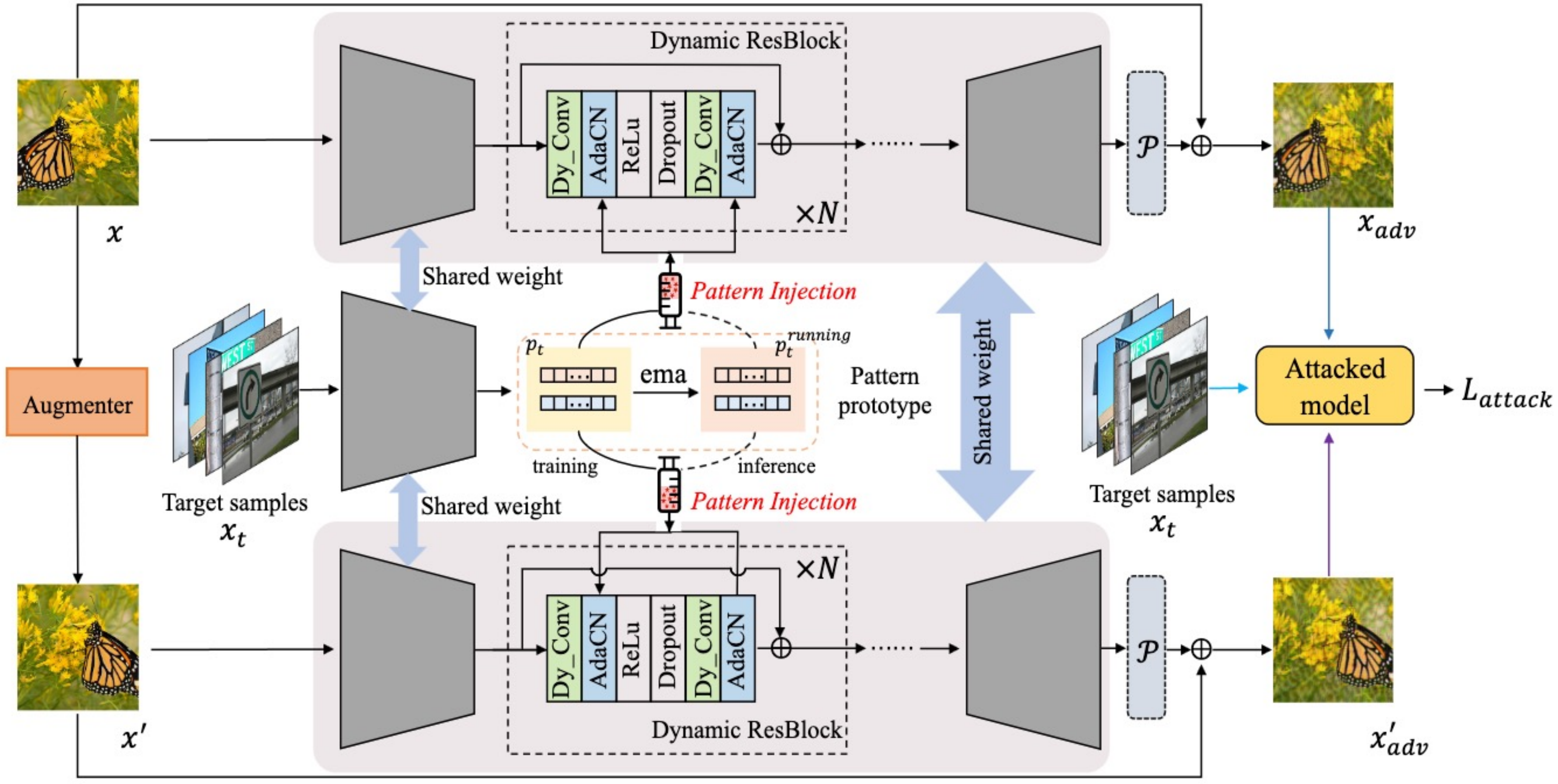
$$P_\theta(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{s \in \mathbb{S}} P_\theta(s \mid \boldsymbol{x}) P_\theta(\boldsymbol{y} \mid \boldsymbol{x}, s). \quad (1)$$

$$P_\theta(\boldsymbol{y}_t \mid \boldsymbol{x}_{adv}) = \sum_{s \in \mathbb{S}} P_\theta(s \mid \boldsymbol{x}_{adv}) P_\theta(\boldsymbol{y}_t \mid \boldsymbol{x}_{adv}, s). \quad (2)$$

- So we propose to exploit $P_\theta\ (y_t \mid x_{adv}, s)$ to perform targeted attacks.
- Injecting the specific style or pattern of images from the given target class y_t can generate targeted transferable adversarial examples.

Method

Architecture of dynamic generative targeted attacks :



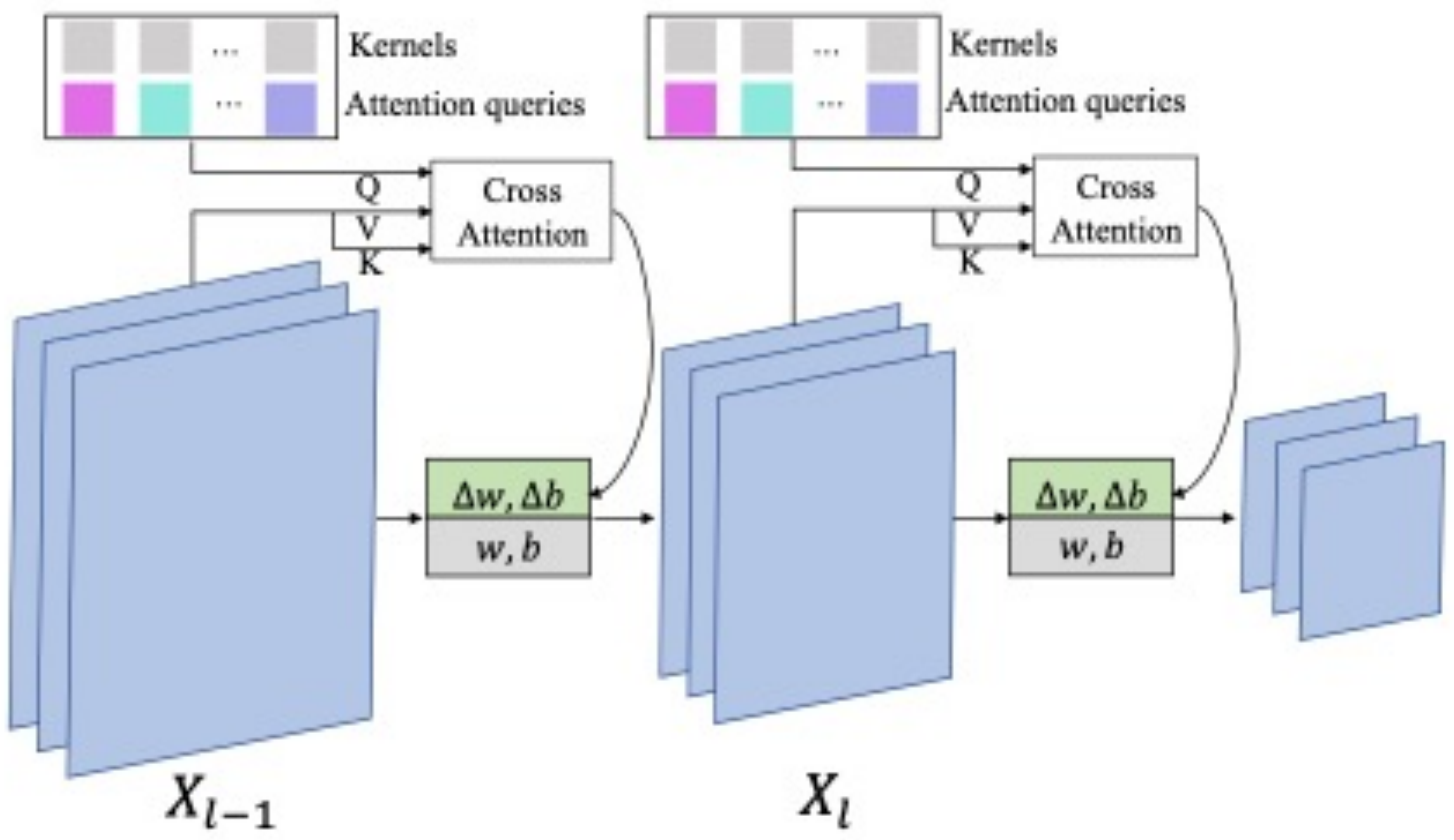
$$\boldsymbol{x}_{adv} = \text{clip}\left\{\text{Proj}\left(\mathcal{W} * G_{\theta(\boldsymbol{x})}(\boldsymbol{x}, p_t), -\varepsilon, \varepsilon\right) + \boldsymbol{x}\right\}, \qquad (3)$$

- **Cross-attention guided dynamic convolution module**
  - **Static convolution**
  - **Dynamic convolution**
- **Pattern injection module:**
  - **Pattern prototype**
  - **Adaptive class norm layer**

Method

➢ **Cross-attention guided dynamic convolution module**

- **Static convolution**
- **Dynamic convolution**



$$\boxed{\boldsymbol{X}_l = \mathrm{conv}(\boldsymbol{X}_{l-1}; W + \Delta W), \quad (4)}$$

$$Q = \boldsymbol{q}W^q,\ K = \boldsymbol{X}_{l-1}^\top W^k \quad V = \boldsymbol{X}_{l-1}^\top W^v$$

$$\Downarrow$$

$$att = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

$$\Downarrow$$

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_N]$$

$$\Downarrow$$

$$\Delta W = \alpha_1 * W_1 + \alpha_2 * W_2 + \cdots + \alpha_N * W_N$$

Benefiting from this dynamic and static mixup convolution operation, our proposed generative attack model can inherit the advantages of both instance-specific and instance-agnostic attacks.

Method

- **Pattern injection module:**
  - **Pattern prototype**
  - **Adaptive class norm layer**

Pattern prototype

$$\boldsymbol{p}_t = \{\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t\}$$

Ema updating

$$\boldsymbol{p}_t^{\text{running}} = \lambda \boldsymbol{p}_t + (1-\lambda)\boldsymbol{p}_t^{\text{running}}$$

Adaptive class norm layer

$$\mathrm{AdaCN}(\boldsymbol{X}) = \boldsymbol{\gamma}_t \left(\frac{\boldsymbol{X} - \mu(\boldsymbol{X})}{\sigma(\boldsymbol{X})}\right) + \boldsymbol{\beta}_t,$$

Theoretical Analyses:

$$\mathcal{X}_s \sim N(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s), \quad \mathcal{X}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t),$$

$$\mathcal{F}_s = \boldsymbol{A}\mathcal{X}_s - \frac{\boldsymbol{A}\boldsymbol{\mu}_s + \boldsymbol{\mu}_t}{2} \sim N(-\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$
$$\mathcal{F}_t = \boldsymbol{E}\mathcal{X}_t - \frac{\boldsymbol{A}\boldsymbol{\mu}_s + \boldsymbol{\mu}_t}{2} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

$$\boldsymbol{\delta} = C_1 \left[\begin{pmatrix} \frac{\sigma_{t1}}{\sigma_{s1}} & & \\ & \ddots & \\ & & \frac{\sigma_{tn}}{\sigma_{sn}} \end{pmatrix} (\boldsymbol{x}_s - \boldsymbol{\mu}_s) + \boldsymbol{\mu}_t \right] - C_2 \boldsymbol{x}_s,$$

Method

Objective functions:

Distance loss

$$\begin{aligned}\mathcal{L} &= D_{KL}\big(f(\boldsymbol{x}_{adv})\|f(\boldsymbol{x}_t)\big) + D_{KL}\big(f(\boldsymbol{x}_t)\|f(\boldsymbol{x}_{adv})\big)\\ \mathcal{L}_{aug} &= D_{KL}\big(f(\boldsymbol{x}'_{adv})\|f(\boldsymbol{x}_t)\big) + D_{KL}\big(f(\boldsymbol{x}_t)\|f(\boldsymbol{x}'_{adv})\big)\end{aligned}$$

Local similarity loss

$$\mathcal{L}_{sim} = \sum_{i,j} \overline{\mathcal{S}}^t_{i,j} \log \frac{\overline{\mathcal{S}}^t_{i,j}}{\overline{\mathcal{S}}_{i,j}} + \sum_{i,j} \overline{\mathcal{S}}_{i,j} \log \frac{\overline{\mathcal{S}}_{i,j}}{\overline{\mathcal{S}}^t_{i,j}},$$

Total objective function

$$\mathcal{L}_{attack} = \mathcal{L} + \mathcal{L}_{aug} + \mathcal{L}_{sim}.$$

Experiments

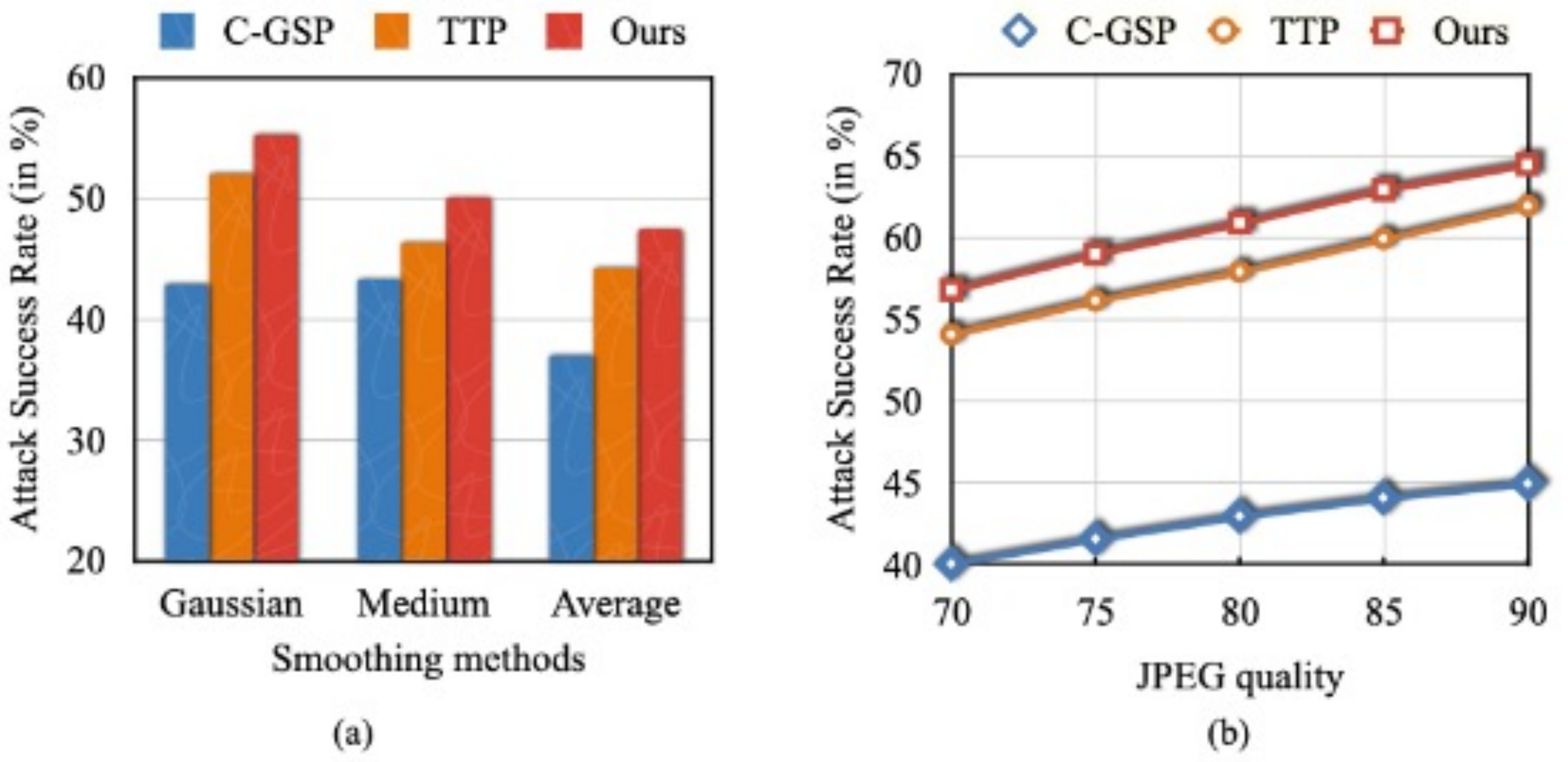
Against normal models

Substite Model	Method	Inc-v3	Inc-v4	Inc-Res-v2	Res152	Densenet-121	GoogleNet	Vgg-16
Inc-v3	MIM	**99.90**	0.80	1.00	0.40	0.20	0.20	0.30
	TI-MIM	98.50	0.50	0.50	0.30	0.20	0.40	0.40
	SI-MIM	99.80	1.50	2.00	0.80	0.70	0.70	0.50
	DIM	95.60	2.70	0.50	0.80	1.10	0.40	0.80
	TI-DIM	96.00	1.10	1.20	0.50	0.50	0.50	0.80
	SI-DIM	90.20	3.80	4.40	2.00	2.20	1.70	1.40
	CD-AP	94.20	57.60	60.10	37.10	41.60	32.30	41.70
	TTP	91.37	46.04	39.37	16.40	33.47	25.80	25.73
	C-GSP	93.40	66.90	**66.60**	41.60	46.40	40.00	45.00
	GAP	86.90	45.06	34.48	34.48	41.74	26.89	34.34
	Ours	94.63	**67.95**	55.03	**50.50**	**47.38**	**47.67**	**48.11**
Res152	MIM	0.50	0.40	0.60	**99.70**	0.30	0.30	0.20
	TI-MIM	0.30	0.30	0.30	96.50	0.30	0.40	0.30
	SI-MIM	1.30	1.20	1.60	99.50	1.00	1.40	0.70
	DIM	2.30	2.20	3.00	92.30	0.20	0.80	0.70
	TI-DIM	0.80	0.70	1.00	90.60	0.60	0.80	0.50
	SI-DIM	4.20	4.80	5.40	90.50	4.20	3.60	2.00
	CD-AP	33.30	43.70	42.70	96.60	53.80	36.60	34.10
	TTP	62.03	49.20	38.70	95.12	82.96	65.09	62.82
	C-GSP	37.70	47.60	45.10	93.20	64.20	41.70	45.90
	GAP	30.99	31.43	20.48	84.86	58.35	29.89	39.70
	Ours	**66.83**	**53.62**	**47.61**	96.48	**86.61**	**68.29**	**69.58**
Vgg-16	MIM	0.26	0.47	0.20	0.35	0.40	0.34	90.24
	TI-MIM	0.43	0.63	0.34	0.55	1.45	0.64	89.13
	SI-MIM	0.35	0.57	0.42	0.31	0.56	0.52	90.89
	DIM	0.75	1.30	0.55	1.00	1.88	1.03	**97.70**
	TI-DIM	0.23	0.38	0.17	0.29	0.48	0.35	93.71
	SI-DIM	0.87	1.12	0.70	0.95	1.89	1.55	91.42
	CD-AP	5.32	8.94	4.87	9.33	14.02	3.19	96.82
	TTP	22.51	17.14	**9.68**	22.68	40.87	22.41	97.59
	C-GSP	9.42	9.60	3.01	11.76	32.28	13.33	96.81
	GAP	3.11	5.26	1.50	5.08	11.23	2.70	93.00
	Ours	**28.18**	**21.78**	9.56	**25.27**	**46.55**	**23.70**	93.00

Experiments

Against robust models and input preprocess defense

Substite Model	Method	Adv-Inc-v3	Ens-IncRes-v2	Res50_SIN	Res50_SIN_IN	Res50_SIN_fine_IN	Res50_AugMix
Res152	MIM	0.19	0.15	0.28	1.58	2.75	0.78
	TI-MIM	0.61	0.73	0.50	2.51	4.75	1.76
	SI-MIM	0.24	0.24	0.39	0.66	0.84	0.36
	DIM	0.63	0.37	0.94	8.50	14.22	3.77
	TI-DIM	0.23	0.30	0.28	0.76	1.49	0.49
	SI-DIM	0.71	0.71	0.75	2.73	3.89	1.37
	CD-AP	3.77	6.48	7.09	63.72	76.79	49.67
	TTP	27.99	26.08	24.61	72.47	74.51	70.96
	GAP	5.72	4.51	7.33	71.04	**83.64**	52.07
	Ours	**31.10**	**30.07**	**27.70**	**77.13**	80.55	**76.78**
VGG16	MIM	0.14	0.15	0.16	0.40	0.34	0.19
	TI-MIM	0.26	0.24	0.20	0.45	0.57	0.28
	SI-MIM	0.28	0.20	0.21	0.49	0.25	0.14
	DIM	0.22	0.16	0.27	0.93	0.99	0.49
	TI-DIM	0.14	0.19	0.21	0.35	0.34	0.21
	SI-DIM	0.50	0.36	0.33	0.80	0.69	0.26
	CD-AP	0.36	0.34	0.35	4.63	10.20	3.60
	TTP	3.75	3.20	**2.66**	27.80	32.70	16.57
	GAP	0.30	0.52	0.42	4.52	8.92	3.35
	Ours	**4.14**	**3.22**	**2.66**	**30.16**	**38.10**	**17.95**



Experiments

Visualization

Summary

- We propose a dynamic generative model to craft transferable targeted adversarial examples with a novel cross-attention guided dynamic convolution module and a pattern injection module.
- Benefit from the dynamic convolution module and a pattern injection module, our method can not only inject pattern or style information of the target class to improve transferable targeted attacks, but also learn specialized convolutional kernels for each input instance, which inherits the advantages of both instance-specific and instance-agnostic attacks.
- We state that *injecting the specific pattern or style of the target class can improve the transferability of targeted adversarial examples*, and we provide a comprehensive theoretical analysis to verify the rationality of this statement.
- We present extensive experiments to demonstrate the effectiveness against normal and robust models.

DEEP SPACE EXPLORATION LAB
深空探测实验室

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thanks for your attention!!!!!!