Content-based Unrestricted Adversarial Attack

Zhaoyu Chen[1,2], Bo Li[2], Shuang Wu[2], Kaixun Jiang[1],
Shouhong Ding[2], Wenqiang Zhang[1,3]
[1]Academy for Engineering and Technology, Fudan University
[2]Youtu Lab, Tencent [3]School of Computer Science, Fudan University

zhaoyuchen20@fudan.edu.cn
https://anonymous.4open.science/r/Adversarial_Content_Attack-3118/

Introduction

- **Background**
  - Existing adversarial defense methods can defend against l_p attacks, but cannot defend against more natural unrestricted attacks.
  - Existing unrestricted attacks are achieved either through reliance on subjective intuition and objective metrics or by implementing minor modifications, thereby constraining their potential for transferability.

- **Challenge**

  We argue that an ideal unrestricted attack should meet three criteria:

  1. maintain the photorealism of the images.
  2. attack content should be diverse, allowing for unrestricted modifications of image contents (shape, texture and color, etc.).
  3. have a high adversarial transferability.

Introduction

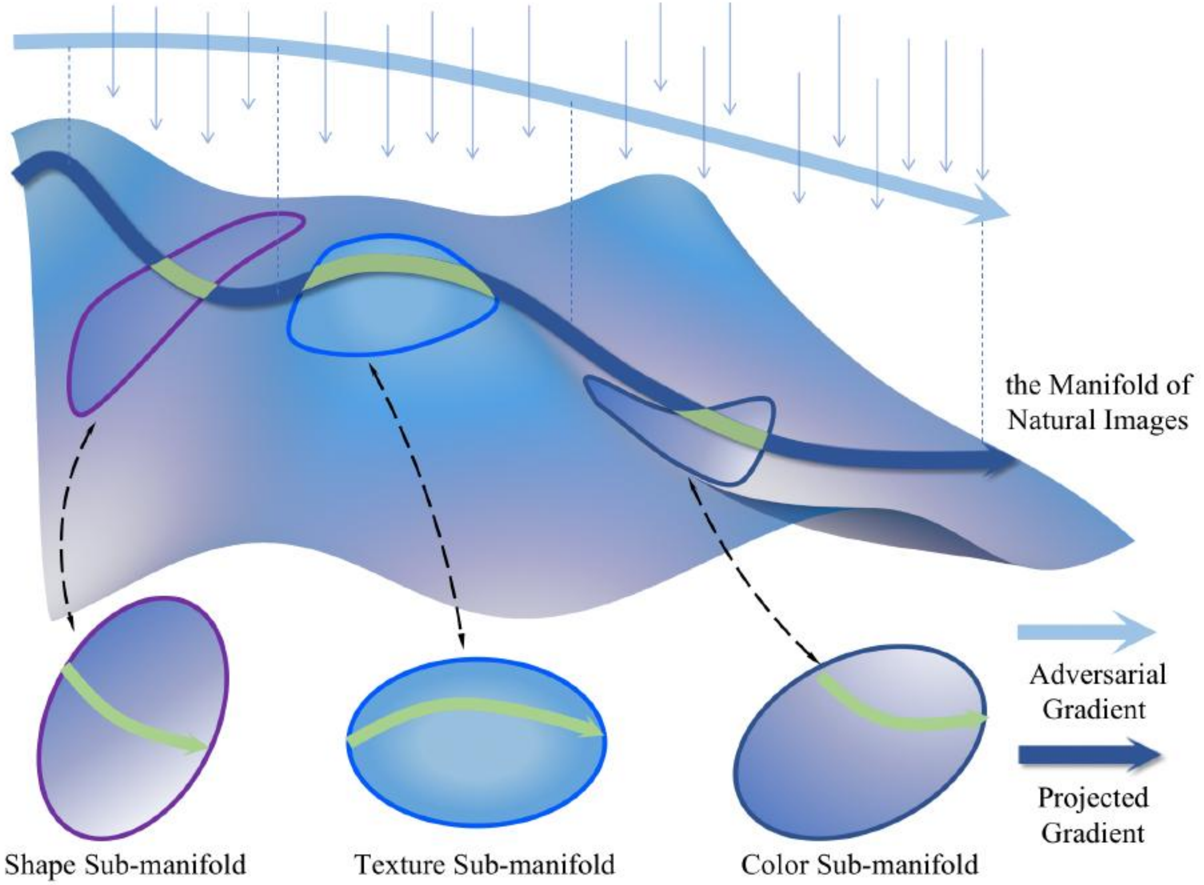
- **Contributions**
  - We propose a novel attack framework called **Content-based Unrestricted Adversarial Attack**, which utilizes high-capacity and well-aligned low-dimensional manifolds to generate adversarial examples that are more diverse and natural in content.
  - We achieve an unrestricted content attack, known as the **Adversarial Content Attack**. By utilizing **Image Latent Mapping** and **Adversarial Latent Optimization** techniques, we optimize latents in a diffusion model, generating high transferable unrestricted adversarial examples.
  - The effectiveness of our attack has been validated through experimentation and visualization. Notably, we have achieved a significant improvement of **13.3~50.4%** over state-of-the-art attacks in terms of adversarial transferability.

Method

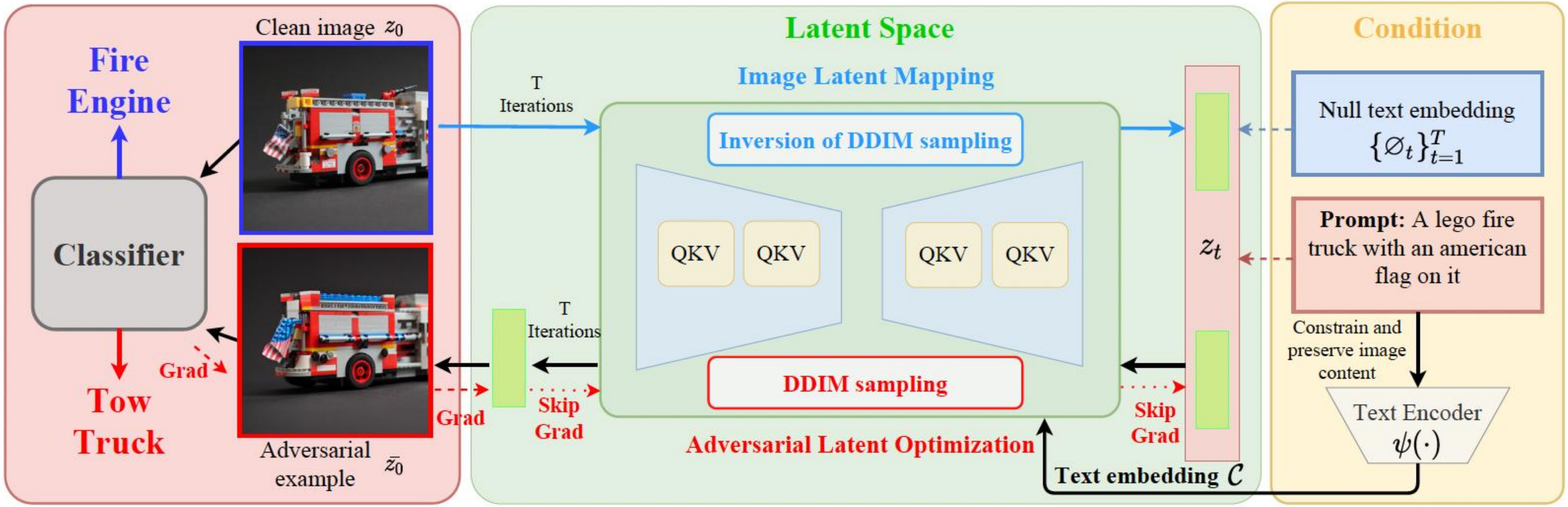
Content-based Unrestricted Adversarial Attack

- We assume that natural images can be mapped onto a low-dimensional manifold by a generative model.
- As this lowdimensional manifold is well-trained on natural images, it naturally ensures the photorealism of the images and possesses the rich content present in natural images.
- Once we map an image onto a low-dimensional manifold, moving it along the adversarial direction on the manifold yields an unrestricted adversarial example.

$$\max_{x_{adv}} \mathcal{L}(\mathcal{F}_\theta(x_{adv}), y), \quad s.t.\ x_{adv} \text{ is natural},$$



Method



Based on the aforementioned framework and the full utilization of the diffusion model's capability, we achieve the unrestricted content-based attack known as **Adversarial Content Attack (ACA)**:

We first employ **Image Latent Mapping (ILM)** to map images onto the latent space represented by this lowdimensional manifold. Subsequently, we introduce an **Adversarial Latent Optimization (ALO)** technique that moves the latent representations of images along the adversarial direction on the manifold. Finally, based on iterative optimization, ACA can generate highly transferable unrestricted adversarial examples that appear quite natural.

Method

- **Image Latent Mapping (ILM)：**

$$\min_{\varnothing_t} \|z^*_{t-1} - z_{t-1}(\bar{z}_t, t, \mathcal{C}, \varnothing_t)\|_2^2, \tag{4}$$

$$z_{t-1}(\bar{z}_t, t, \mathcal{C}, \varnothing_t) = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}\bar{z}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \tilde{\epsilon}_\theta(z_t, t, \mathcal{C}, \varnothing_t). \tag{5}$$

- **Adversarial Latent Optimization (ALO):**

$$g_k \leftarrow \mu \cdot g_{k-1} + \frac{\nabla_{z_T}\mathcal{L}\left(\mathcal{F}_\theta\left((\varrho(\bar{z}_0), y\right)\right)}{\|\nabla_{z_T}\mathcal{L}\left(\mathcal{F}_\theta\left(\varrho(\bar{z}_0), y\right)\right)\|_1}, \tag{12}$$

$$\delta_k \leftarrow \Pi_\kappa\left(\delta_{k-1} + \eta \cdot \mathrm{sign}(g_k)\right). \tag{13}$$

  - **Skip Gradient**

$$\nabla_{z_T}\mathcal{L}\left(\mathcal{F}_\theta(\bar{z}_0), y\right) = \rho\frac{\partial\mathcal{L}}{\partial\bar{z}_0}$$

  - **Differentiable Boundary Processing**

$$\varrho(x) = \begin{cases} \tanh(1000x)/10000, & x < 0, \\ x, & 0 \le x \le 1, \\ \tanh(1000(x-1))/10001, & x > 1. \end{cases} \tag{11}$$

Algorithm 1 Adversarial Content Attack

Input: a input image z_0 with the label y, a text embedding $\mathcal{C} = \psi(\mathcal{P})$, a classifier $\mathcal{F}_\theta(\cdot)$, DDIM steps T, image mapping iteration N_i, attack iterations N_a, and momentum factor μ

```
Calculate latents {z_0^*, ..., z_T^*} using Equation 5 over z_0 with w = 1
Initialize w = 7.5, z̄_T ← z_T^*, ∅ ← ψ(""), δ_0 ← 0, g_0 ← 0
// Image Latent Mapping
for t = T, T − 1 ..., 1 do
    for j = 1, ..., N_i do
        ∅_t ← ∅_t − ζ∇_{∅_t} ||z_{t−1}^* − z_{t−1}(z̄_t, t, C, ∅_t)||_2^2
    end for
    z̄_{t−1} ← z_{t−1}(z̄_t, t, C, ∅_t), ∅_{t−1} ← ∅_t
end for
// Adversarial Latent Optimization
for k = 1, ..., N_a do
    z̄_0 ← Ω(z̄_T + δ_{k−1}, T, C, {∅_t}_{t=1}^T)
    g_k ← μ · g_{k−1} + ∇_{z_T}L(F_θ(ϱ(z̄_0), y)) / ||∇_{z_T}L(F_θ(ϱ(z̄_0), y))||_1
    δ_k ← Π_κ(δ_{k−1} + η · sign(g_k))
end for
z̄_0 ← ϱ(Ω(z̄_T + δ_{N_a}, T, C, {∅_t}_{t=1}^T))
```

Output: The unrestricted adversarial example $\bar{z}_0$.

Experiments

- **Dataset**
  ImageNet-compatible Dataset

- **Metric**
  Attack Success Rate (ASR)

- **State-of-the-art methods**
  - SAE
  - ADef
  - ReColorAdv
  - cAdv
  - tAdv
  - ColorFool
  - NCF

Table 1: Performance comparison of adversarial transferability on normally trained CNNs and ViTs. We report attack success rates (%) of each method ("*" means white-box attack results).

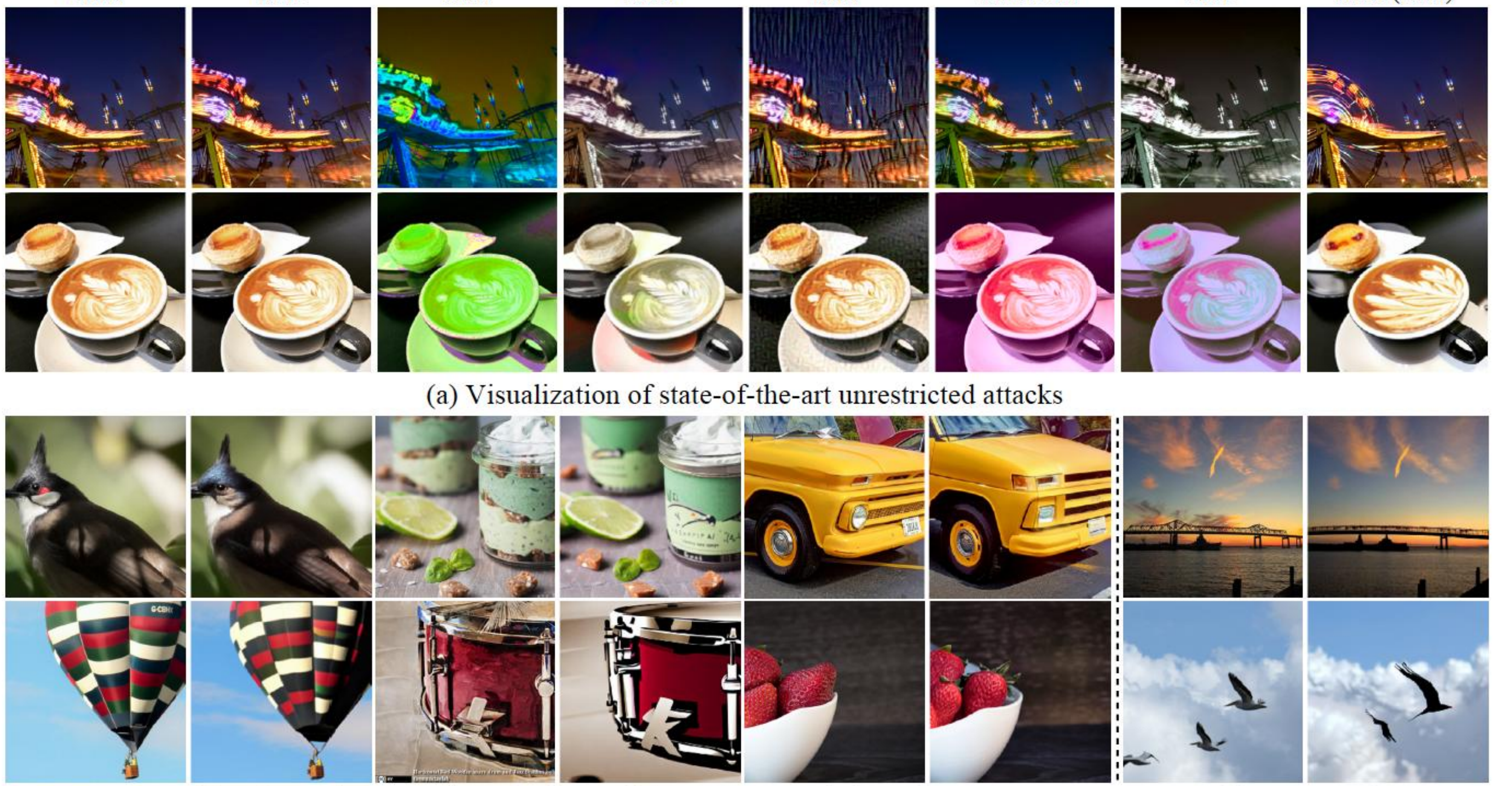
Surrogate Model	Attack	Models										Avg. ASR (%)
		CNNs						Transformers				
		MN-v2	Inc-v3	RN-50	Dense-161	RN-152	EF-b7	MobViT-s	ViT-B	Swin-B	PVT-v2	
-	Clean	12.1	4.8	7.0	6.3	5.6	8.7	7.8	8.9	3.5	3.6	6.83
	ILM	13.5	5.5	8.0	6.3	5.9	8.3	8.3	9.0	4.8	4.0	7.36
MobViT-s	SAE	60.2	21.2	54.6	42.7	44.9	30.2	82.5*	38.6	21.1	20.2	37.08
	ADef	14.5	6.6	9.0	8.0	7.1	9.8	80.8*	9.7	5.1	4.6	8.27
	ReColorAdv	37.4	14.7	26.7	22.4	21.0	20.8	96.1*	21.5	16.3	16.7	21.94
	cAdv	41.9	25.4	33.2	31.2	28.2	34.7	84.3*	32.6	22.7	22.0	30.21
	tAdv	33.6	18.8	22.1	18.7	18.7	15.8	**97.4***	15.3	11.2	13.7	18.66
	ACE	30.7	9.7	20.3	16.3	14.4	13.8	99.2*	16.5	6.8	5.8	14.92
	ColorFool	47.1	12.0	40.0	28.1	30.7	19.3	81.7*	24.3	9.7	10.0	24.58
	NCF	**67.7**	31.2	60.3	41.8	52.2	32.2	74.5*	39.1	20.8	23.1	40.93
	ACA (Ours)	66.2	**56.6**	**60.6**	**58.1**	**55.9**	**55.5**	89.8*	**51.4**	**52.7**	**55.1**	**56.90**
MN-v2	SAE	90.8*	22.5	53.2	38.0	41.9	26.9	44.6	33.6	16.8	18.3	32.87
	ADer	56.6*	7.6	8.4	7.7	7.1	10.9	11.7	9.5	4.5	4.5	7.99
	ReColorAdv	97.7*	18.6	33.7	24.7	26.4	20.7	31.8	17.7	12.2	12.6	22.04
	cAdv	96.6*	26.8	39.6	33.9	29.9	32.7	41.9	33.1	20.6	19.7	30.91
	tAdv	**99.9***	27.2	31.5	24.3	24.5	22.4	40.5	16.1	15.9	15.1	24.17
	ACE	99.1*	9.5	17.9	12.4	12.6	11.7	16.3	12.1	5.4	5.6	11.50
	ColorFool	93.3*	9.5	25.7	15.3	15.4	13.4	15.7	14.2	5.9	6.4	13.50
	NCF	93.2*	33.6	**65.9**	43.5	**56.3**	33.0	52.6	35.8	21.2	20.6	40.28
	ACA (Ours)	93.1*	**56.8**	62.6	**55.7**	56.0	**51.0**	**59.6**	**48.7**	**48.6**	**50.4**	**54.38**
RN-50	SAE	63.2	25.9	88.0*	41.9	46.5	28.8	45.9	35.3	20.3	19.6	36.38
	ADer	15.5	7.7	55.7*	8.4	7.8	11.4	12.3	9.2	4.6	4.9	9.09
	ReColorAdv	40.6	17.7	96.4*	28.3	33.3	19.2	29.3	18.8	12.9	13.4	23.72
	cAdv	44.2	25.3	97.2*	36.8	37.0	34.9	40.1	30.6	19.3	20.2	32.04
	tAdv	43.4	27.0	**99.0***	28.8	30.2	21.6	35.9	16.5	15.2	15.1	25.97
	ACE	32.8	9.4	99.1*	16.1	15.2	12.7	20.5	13.1	6.1	5.3	14.58
	ColorFool	41.6	9.8	90.1*	18.6	21.0	15.4	20.4	15.4	5.9	6.8	17.21
	NCF	**71.2**	33.6	91.4*	48.5	60.5	32.4	52.6	36.8	19.8	21.7	41.90
	ACA (Ours)	69.3	**61.6**	88.3*	**61.9**	**61.7**	**60.3**	**62.6**	**52.9**	**51.9**	**53.2**	**59.49**
ViT-B	SAE	54.5	26.9	49.7	38.4	41.4	30.4	46.1	78.4*	19.9	18.1	36.16
	ADer	15.3	8.3	9.9	8.4	7.6	12.0	12.4	81.5*	5.3	5.5	9.41
	ReColorAdv	25.5	12.1	17.5	13.9	14.4	15.4	22.9	**97.7***	10.9	8.6	15.69
	cAdv	31.4	27.0	26.1	22.5	19.9	26.1	32.9	96.5*	18.4	16.9	24.58
	tAdv	39.5	22.8	25.8	23.2	22.3	20.8	34.1	93.5*	16.3	15.3	24.46
	ACE	30.9	11.4	22.0	15.5	15.2	13.0	17.0	98.6*	6.5	6.3	15.31
	ColorFool	45.3	13.9	35.7	24.3	28.8	19.8	27.0	83.1*	8.9	9.3	23.67
	NCF	55.9	25.3	50.6	34.8	42.3	29.9	40.6	81.0*	20.0	19.1	35.39
	ACA (Ours)	**64.6**	**58.8**	**60.2**	**58.1**	**58.1**	**57.1**	**60.8**	87.7*	**55.5**	**54.9**	**58.68**

Experiments

Table 2: Performance comparison of adversarial transferability on adversarial defense methods.

Attack	HGD	R&P	NIPS-r3	JPEG	Bit-Red	DiffPure	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	Res-De	Shape-Res	Avg. ASR (%)
Clean	1.2	1.8	3.2	6.2	17.6	15.4	6.8	8.9	2.6	4.1	6.7	6.77
ILM	1.5	1.9	3.5	7.1	18.5	16.1	6.8	9.8	3.0	5.1	8.1	7.40
SAE	21.4	19.0	25.2	25.7	43.5	39.8	25.7	29.6	20.0	35.1	49.6	30.42
ADer	2.9	3.6	6.9	10.4	27.5	18.1	10.1	12.1	5.6	6.0	9.7	10.26
ReColorAdv	5.1	7.0	10.0	20.0	24.3	20.0	11.1	15.5	7.4	11.6	18.4	13.67
cAdv	12.2	14.0	17.7	11.1	33.9	32.9	19.9	23.2	14.6	16.2	25.3	20.09
tAdv	10.9	12.4	14.4	17.8	29.6	21.2	17.7	19.0	12.5	16.4	25.4	17.94
ACE	4.9	5.9	11.1	12.6	28.1	24.9	12.4	15.4	7.6	11.6	21.0	14.14
ColorFool	9.1	9.6	15.3	18.0	37.9	33.8	17.8	21.3	10.5	20.3	35.0	20.78
NCF	22.8	21.1	25.8	26.8	43.9	39.6	27.4	31.9	21.8	34.4	47.5	31.18
ACA (Ours)	**52.2**	**53.6**	**53.9**	**59.7**	**63.4**	**63.7**	**59.8**	**62.2**	**53.6**	**55.6**	**60.8**	**58.05**

Experiments



(a) Visualization of state-of-the-art unrestricted attacks

(b) Adversarial examples of Adversarial Content Attack (ACA)

(c) Case Study

Table 3: Image quality assessment.

Attack	NIMA -AVA↑	HyperIQA↑	MUSIQ -AVA↑	MUSIQ -KonIQ↑	TReS↑
Clean	5.15	0.667	4.07	52.66	82.01
ILM	5.15	0.672	4.08	52.55	81.80
SAE	5.05	0.597	3.79	47.24	71.88
ADer	4.89	0.608	3.89	47.39	72.10
ReColorAdv	5.07	0.668	3.97	51.08	80.32
cAdv	4.97	0.623	3.87	48.32	75.12
tAdv	4.83	0.525	3.78	44.71	67.07
ACE	5.12	0.648	3.96	50.49	77.25
ColorFool	5.24	0.662	4.05	52.27	78.54
NCF	4.96	0.634	3.87	50.33	74.10
ACA (Ours)	**5.54**	**0.691**	**4.37**	**56.08**	**85.11**

Experiments

Table 4: Attack speed of unrestricted attacks. We choose MN-v2 as the surrogate model and evaluate the inference time on an NVIDIA Tesla A100.

Attack	SAE	ADer	ReColorAdv	cAdv	tAdv	ACE	ColorFool	NCF	ACA (Ours)
Time (sec)	8.80	0.41	3.86	18.67	4.88	6.64	12.18	10.45	60.0+65.33=125.33

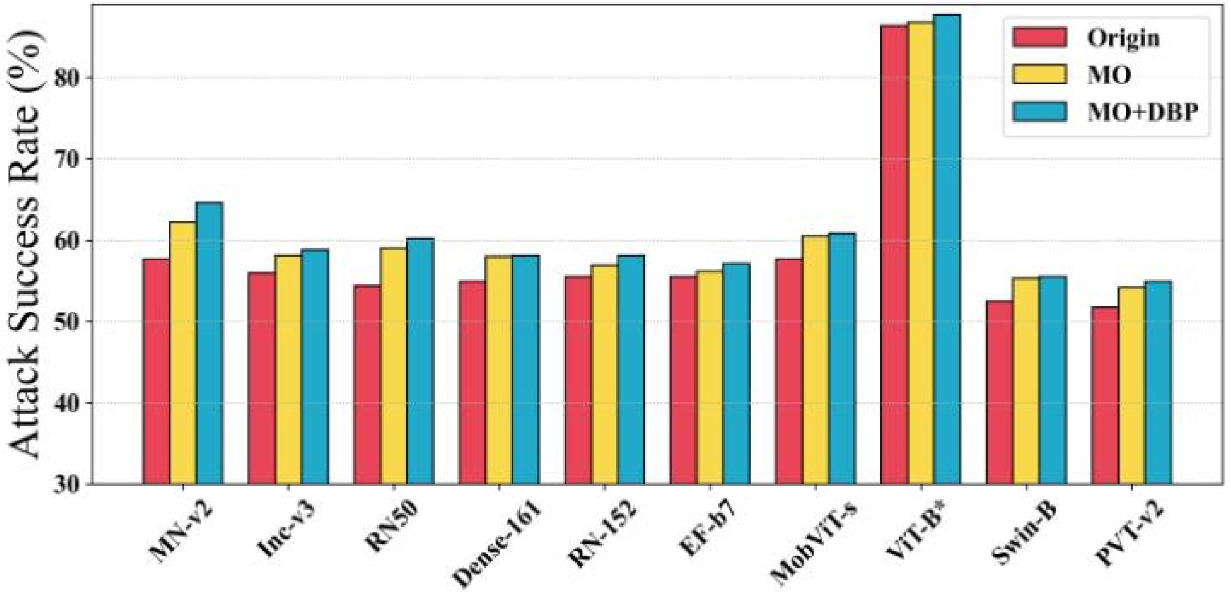


Figure 4: Ablation studies of momentum (MO) and differentiable boundary processing (DBP).


THANK
YOU