ICLR

University of Chinese Academy of Sciences

Resolution Attack: Exploiting Image Compression to Deceive Deep Neural Networks

Wangjia Yu[1,2], **Xiaomeng Fu**[1,2], **Qiao Li**[1,2], **Jizhong Han**[1], **Xiaodan Zhang**[1*]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{yuwangjia,fuxiaomeng,liqiao,hanjizhong,zhangxiaodan}@iie.ac.cn

code

Agenda

- Background
- Introduction
- Method
- Experimental results
- Conclusion

Background

Model Robustness: Essential for maintaining performance under uncertainties like input noise and adversarial attacks.

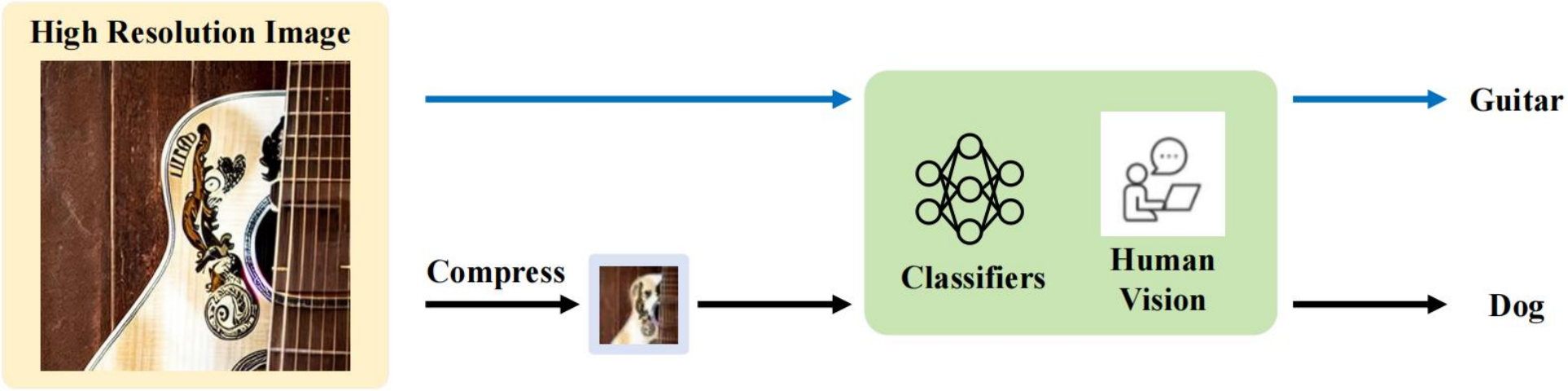
Research Gap: Limited exploration of robustness across different image resolutions.

Real-World Relevance: Classifiers encounter varying image resolutions due to compression and capture limitations.

Objective: Introduce resolution attacks to assess and highlight vulnerabilities of classifiers to resolution variations.

Introduction

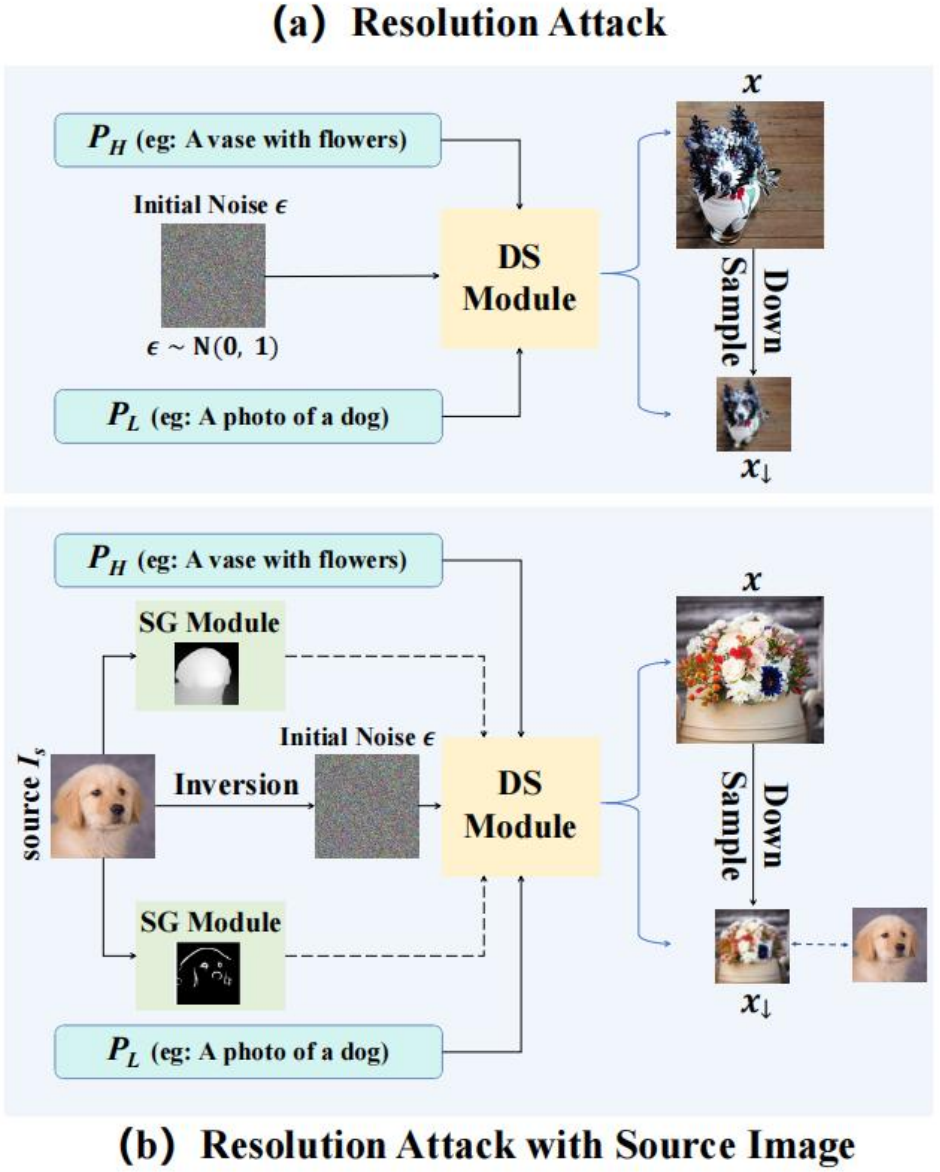
Resolution attacks generate images with **dual semantic representations**, where a high-resolution image may be misclassified after compression. This highlights the vulnerability of classifiers to resolution changes and offers new tools for assessing and enhancing model robustness.



Method

We categorize the RA attack into two types:

- Resolution Attack: Generates images with dual semantic representations across resolutions without additional constraints.
- Resolution Attack with Source Image: Involves manipulating a specific source image to achieve dual representations across resolutions, with supplementary constraints.

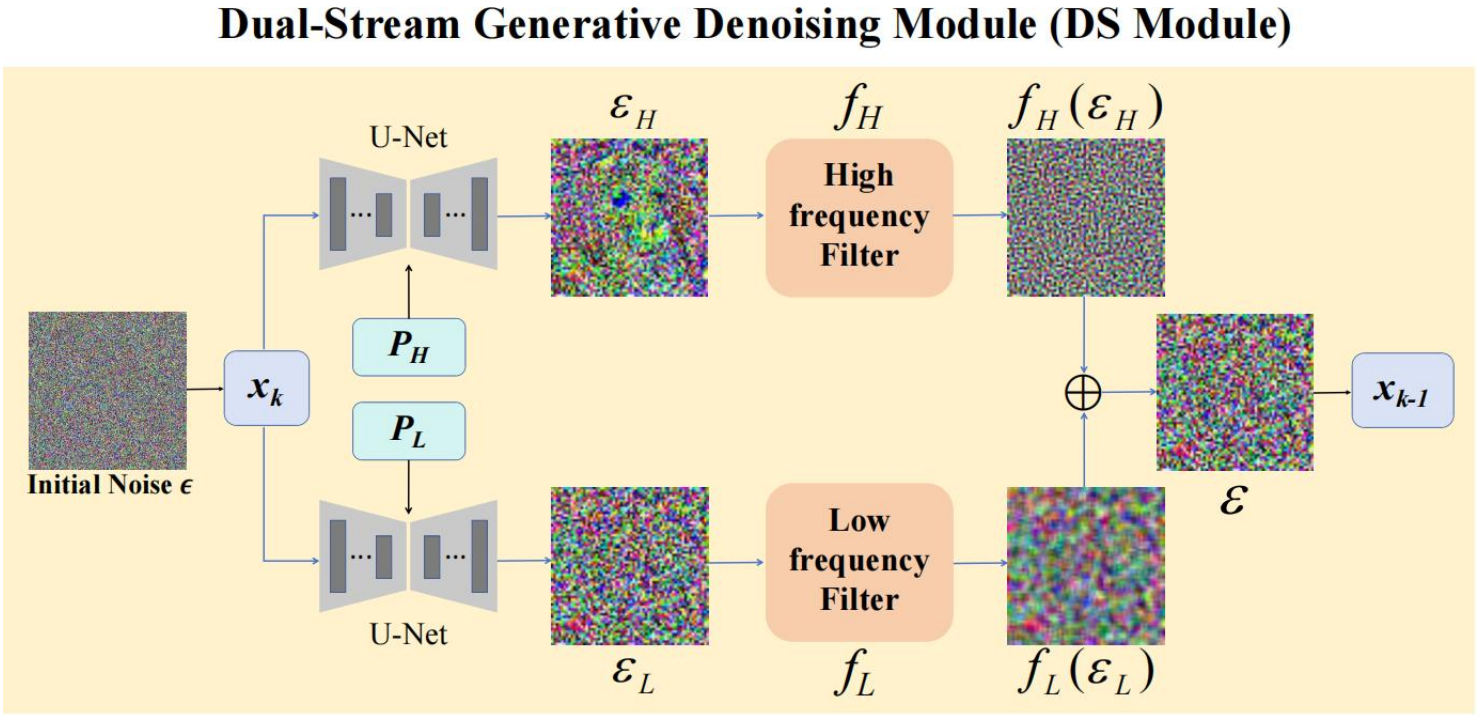


Method - Resolution Attack (RA)

Dual-Stream Generative Denoising Module: Uses diffusion models to generate high-resolution images with dual semantics.

Time-Dependent Denoising Strategy: Three-stage process to ensure smooth transition between resolutions.

- Early Stage: Focus on low-resolution semantics (C_L).
- Middle Stage: Combine both low and high-resolution semantics.
- Late Stage: Refine high-resolution details (C_H).

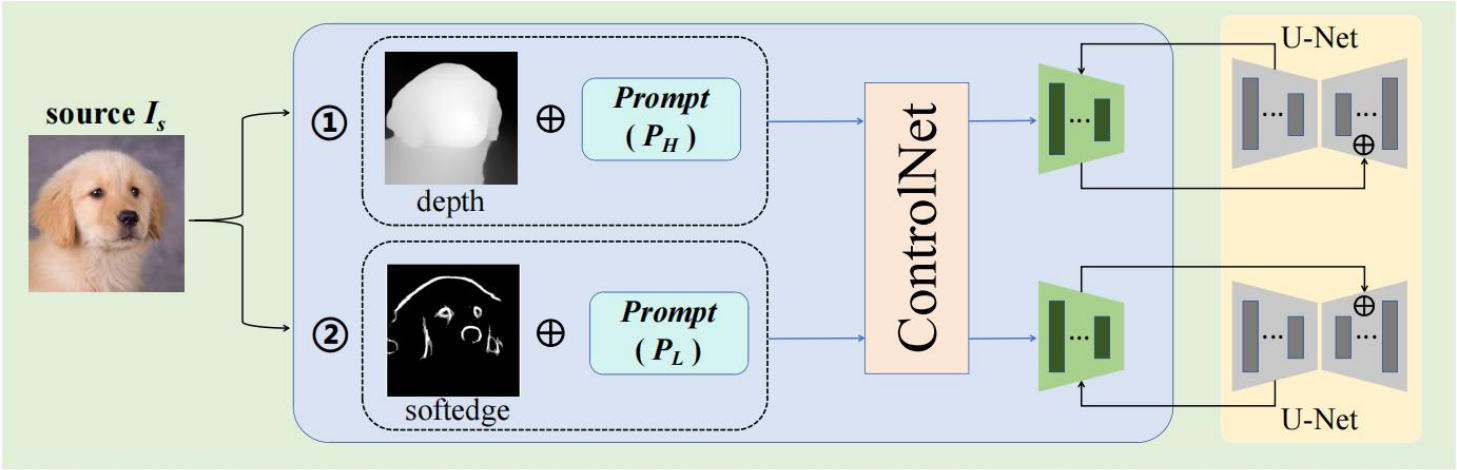


Method - Resolution Attack with Source Image (RAS)

DDIM Inversion: Maps the source image to initial noise to preserve structural attributes.

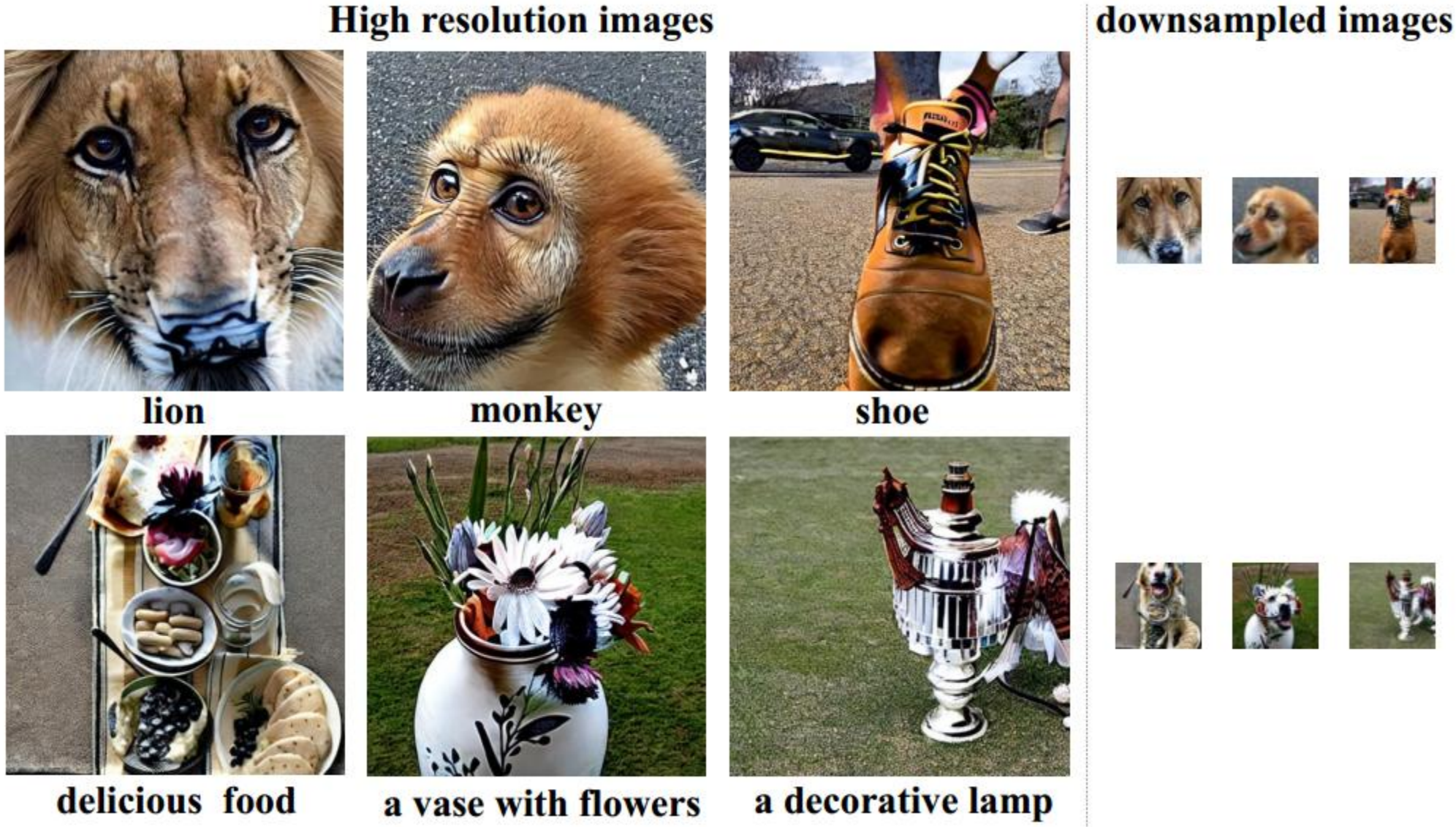
Structural Guidance Module: Integrates depth maps and softedge maps to maintain source image structure.

- Depth Maps: Use depth maps on high-resolution prompts (P_H) to provide rough structural guidance.
- Softedge Maps: Use softedge maps on low-resolution prompts (P_L) to provide precise shape guidance.

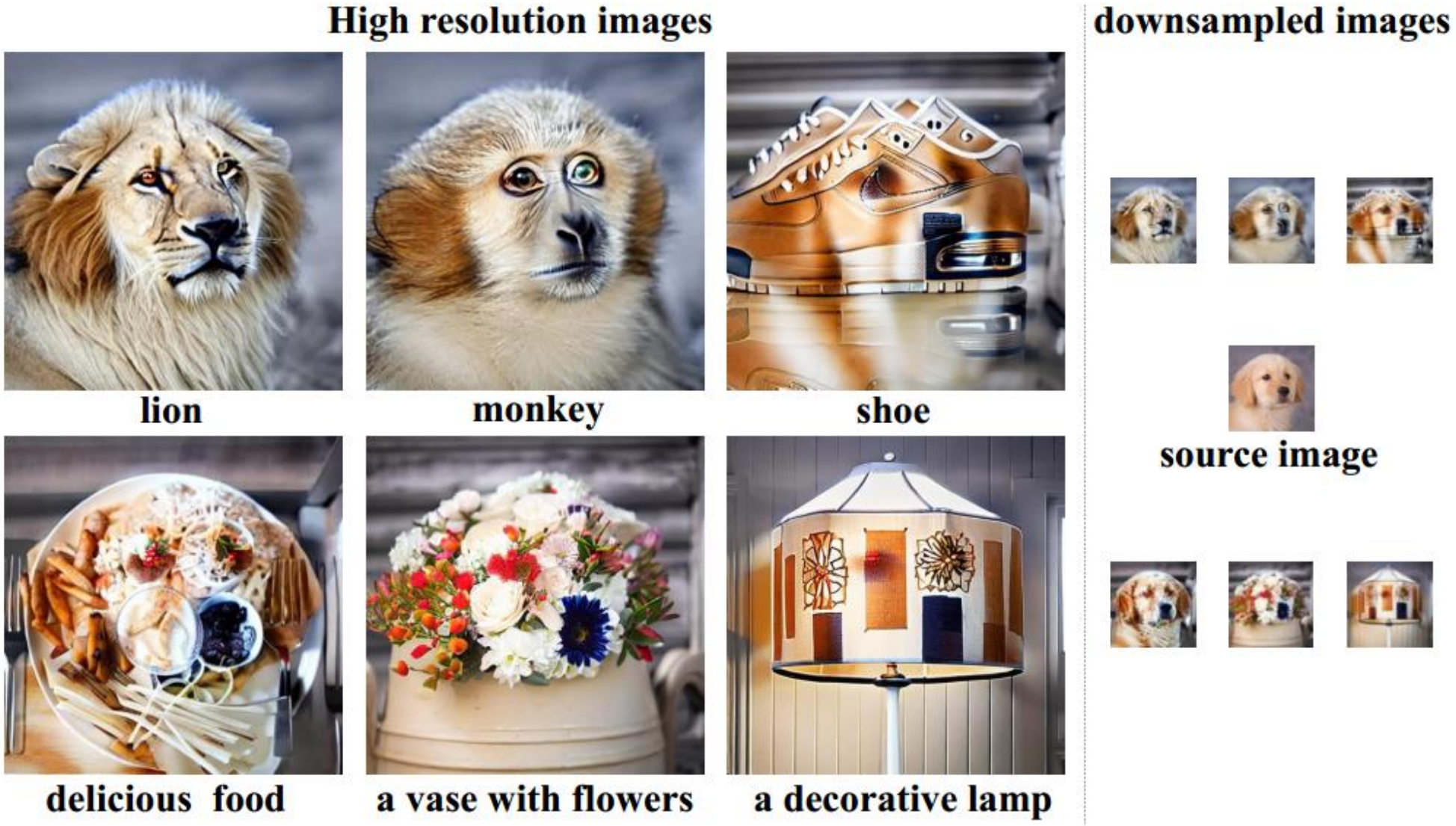


Structural Guidance Module (SG Module)

Experiment - Qualitative Results (RA)



Experiment - Qualitative Results (RAS)



Experiment - Quantitative Results

Table 1: The quantitative results of the Resolution Attack.

Classifiers	Labeled Attack					Unlabeled Attack		
	$Acc_H \uparrow$	$Acc_L \uparrow$	$ASR_C \uparrow$	$CLIP_H \uparrow$	$CLIP_L \uparrow$	$Acc_L \uparrow$	$CLIP_H \uparrow$	$CLIP_L \uparrow$
Resnet-50	68.5%	71.8%	71.4%			63.2%		
VGG19	65.3%	64.8%	61.6%			59.5%		
InceptionV3	76.7%	43.3%	43.6%	0.298	0.248	26.0%	0.256	0.247
EfficientNet	89.6%	67.3%	68.2%			42.0%		
DenseNet	69.8%	63.4%	60.5%			69.3%		

Table 2: The quantitative results of the Resolution Attack with Source image.

Classifiers	Labeled Attack						Unlabeled Attack			
	$Acc_H \uparrow$	$Acc_L \uparrow$	$ASR_C \uparrow$	$CLIP_H \uparrow$	$CLIP_L \uparrow$	SSIM↑	$Acc_L \uparrow$	$CLIP_H \uparrow$	$CLIP_L \uparrow$	SSIM↑
Resnet-50	59.5%	78.6%	77.1%				38.8%			
VGG19	58.1%	77.1%	74.9%				40.0%			
InceptionV3	61.3%	46.0%	42.6%	0.295	0.247	0.727	18.9%	0.266	0.217	0.660
EfficientNet	79.9%	70.7%	70.3%				30.5%			
DenseNet	57.4%	77.5%	75.3%				44.0%			

Tables 1 and 2 present the quantitative results of RA and RAS on convolutional neural network-based classifiers. The experimental results indicate that all classifiers are susceptible to resolution attacks.

Experiment - Quantitative Results

To assess the effectiveness of attacks across different model architectures, experiments were conducted on seven models from three different categories (ViT, Feature Pyramid, VLMs).

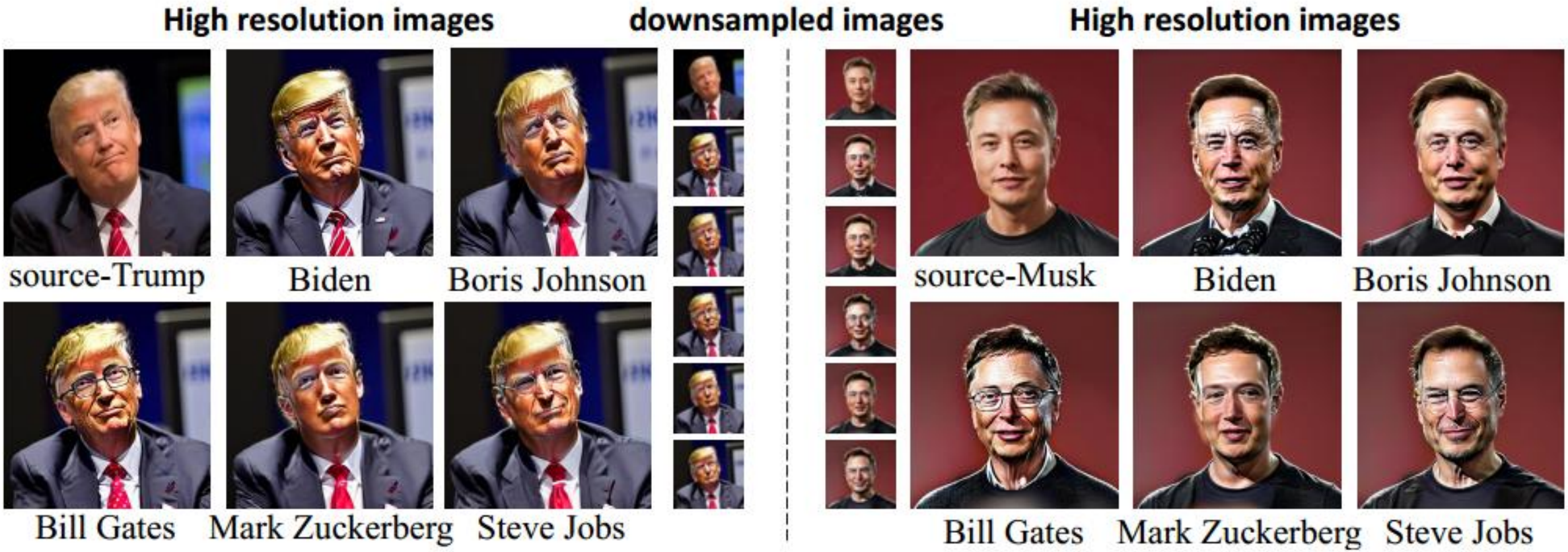
Table 4: Additional quantitative results of the Resolution Attack.

Classifiers	Labeled Attack			Unlabeled Attack
	$\text{Acc}_\text{H} \uparrow$	$\text{Acc}_\text{L} \uparrow$	$\text{ASR}_\text{C} \uparrow$	$\text{Acc}_\text{L} \uparrow$
ViT-b32	55.7%	58.1%	41.3%	83.2%
ViT-l32	50.3%	65.3%	44.1%	87.2%
Fasterrcnn_resnet50_fpn	——	47.7%	——	19.4%
Maskrcnn_resnet50_fpn	——	57.5%	——	29.1%
Clip(zero-shot)	92.1%	34.9%	33.3%	62.8%
Blip2(image caption)	72.1%	58.7%	53.7%	91.2%
LLAVA(VQA)	71.8%	68.9%	63.6%	90.0%

Table 5: Additional quantitative results of the Resolution Attack with Source image.

Classifiers	Labeled Attack			Unlabeled Attack
	$\text{Acc}_\text{H} \uparrow$	$\text{Acc}_\text{L} \uparrow$	$\text{ASR}_\text{C} \uparrow$	$\text{Acc}_\text{L} \uparrow$
ViT-b32	44.9%	74.5%	56.6%	58.0%
ViT-l32	36.9%	83.9%	60.7%	62.4%
Fasterrcnn_resnet50_fpn	——	53.5%	——	11.8%
Maskrcnn_resnet50_fpn	——	67.8%	——	18.6%
Clip(zero-shot)	85.7%	44.7%	39.9%	30.9%
Blip2(image caption)	61.4%	63.5%	55.5%	51.8%
LLAVA(VQA)	63.8%	71.6%	59.7%	51.4%

Experiment - Resolution attacks as the face swapper



Conclusion

Contributions: Defined resolution attacks and developed frameworks for RA and RAS attack.

Implications: Revealed vulnerabilities of current classifiers to resolution variations.

Future Work: Explore defense mechanisms against resolution attacks and further applications of the proposed frameworks.