2021 WEST LAKE CYBERSECURITY CONFERENCE
西湖论剑·网络安全大会

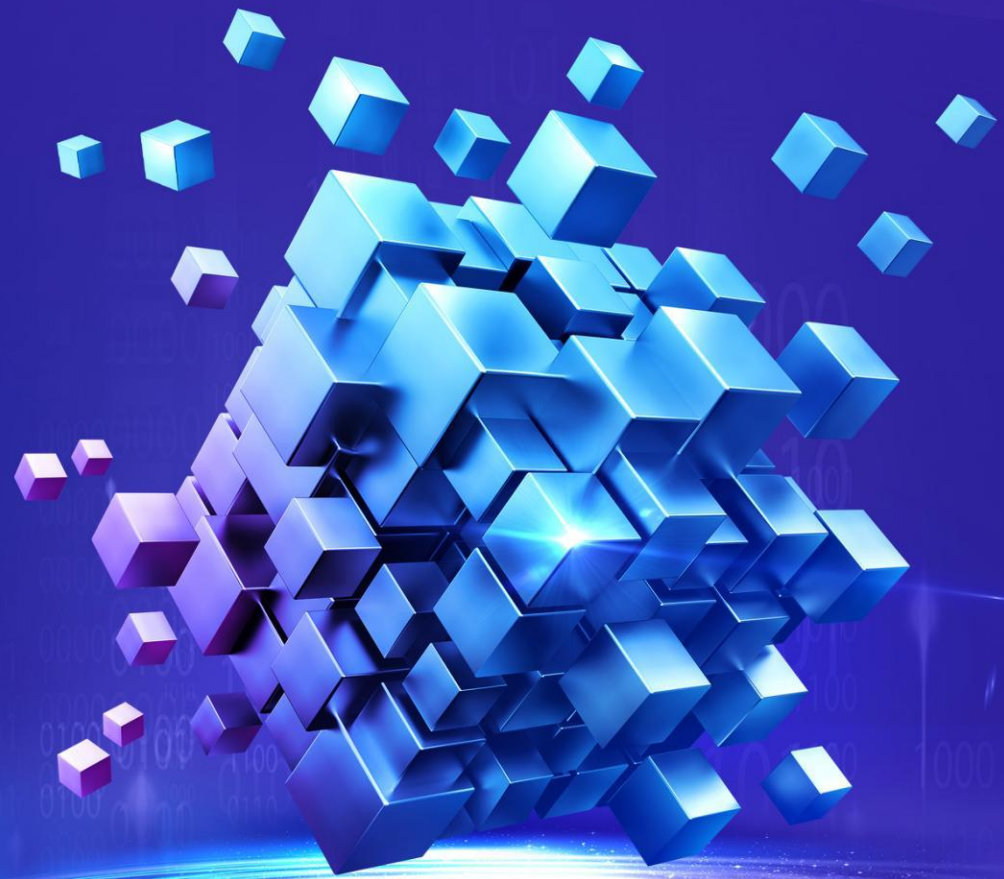
2021 CYBERSECURITY : THE FOUNDATION OF DIGITAL REFORM

让分享更安全

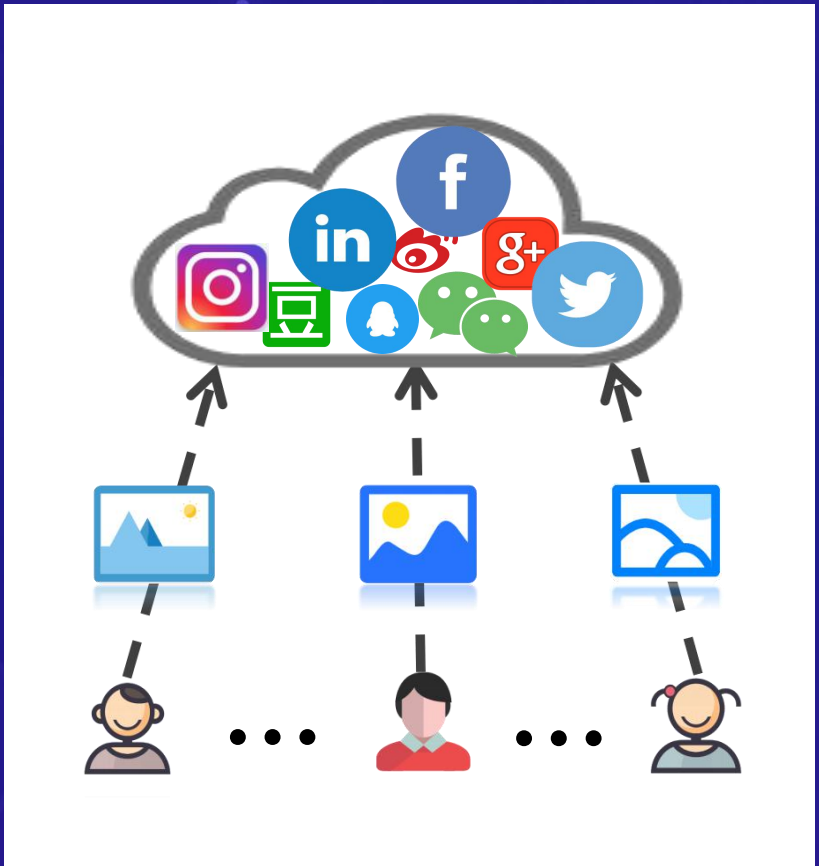
——抗压缩“隐形衣”

演讲人：王志波 教授

浙江大学网安/计算机学院



图像分享已成为在线社交网络的流行特质



- Facebook上每天大约有3亿张照片上传[1]
- Instagram上每天有超过9500万张照片上传，到目前为止，已经有超过400亿张照片被分享[2]
- 研究显示，每天有32亿张图片被分享[2]

[1] https://zephoria.com/social-media/top-15-valuable-facebook-statistics/
[2] https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/

图像包含丰富个人信息，泄露隐私风险大

爬取&分析

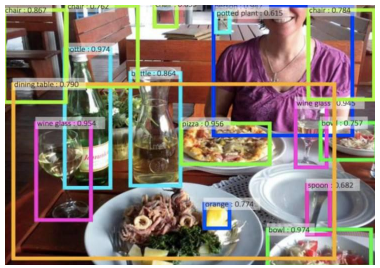
隐私信息：位置、
人际关系、喜好、...

- 定向广告
- 市场营销
- ...
- 经济诈骗

Clearview AI公司从互联网上（Facebook、YouTube和Venmo等）爬取了超过30亿张图片以构建他们的面部识别系统，该系统在没有公众监督的情况下已提供给了600家机构进行使用

基于深度学习的图像识别技术加剧了图像隐私泄露的风险

图像分类



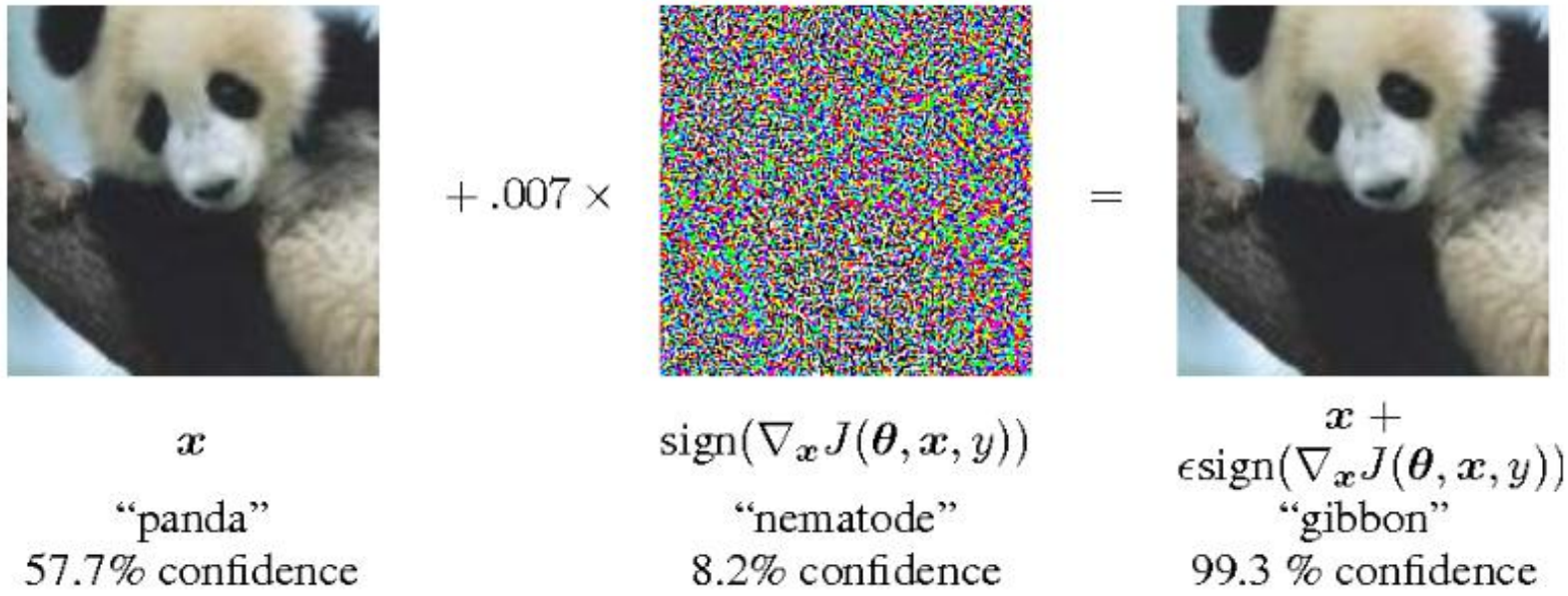
目标检测

人脸检测与分析

图像字幕生成

- 一方面，各种各样的深度神经网络可以从大量的图像中自动提取丰富的信息
- 另一方面，用于隐私保护的传统模糊处理技术（如人脸模糊和马赛克）无法抵御基于DNNs的图像识别模型

深度神经网络易受对抗样本的误导

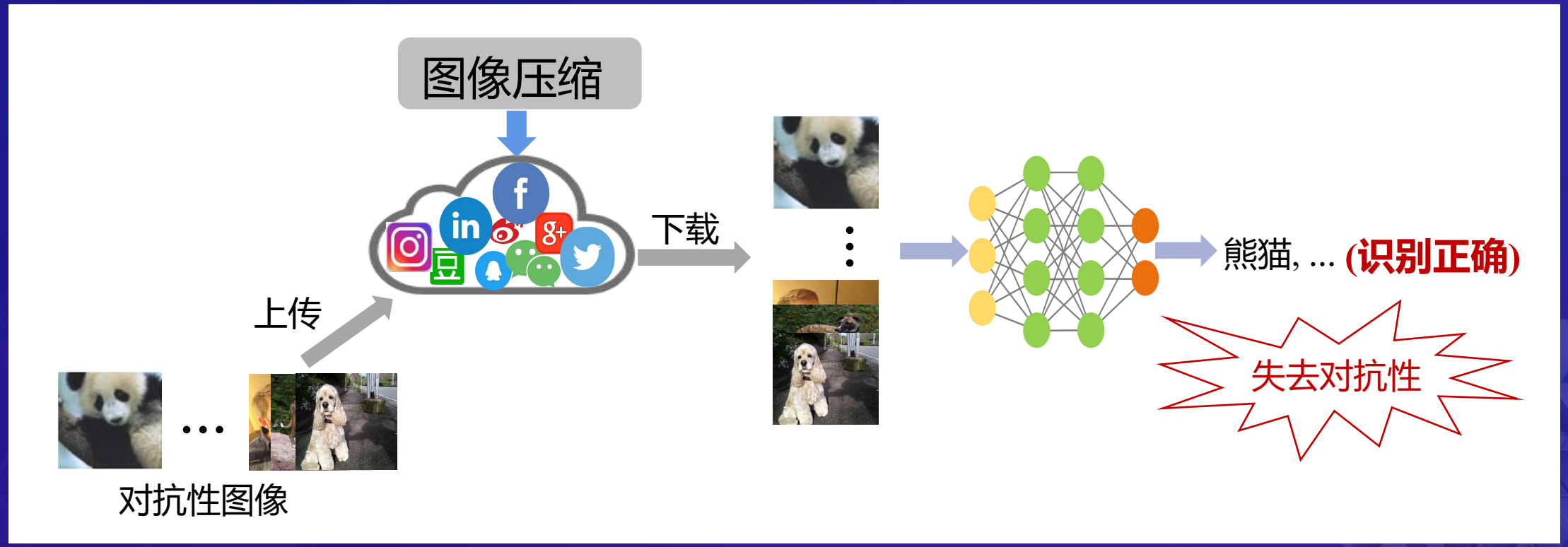


相比于传统的模糊处理技术，对抗样本可以在不影响图像视觉效果的情况下误导DNNs，更适用于保护社交网络上图像的隐私。

[1] Ian J Goodfellow, et al. Explaining and Harnessing Adversarial Examples (arXiv'14).

对抗性图像经过有损压缩后易失去对抗性

- 图像压缩技术广泛应用于社交平台来节约通信资源和提高访问效率，但其能破坏精心构造的对抗性扰动，导致对抗性图像失效

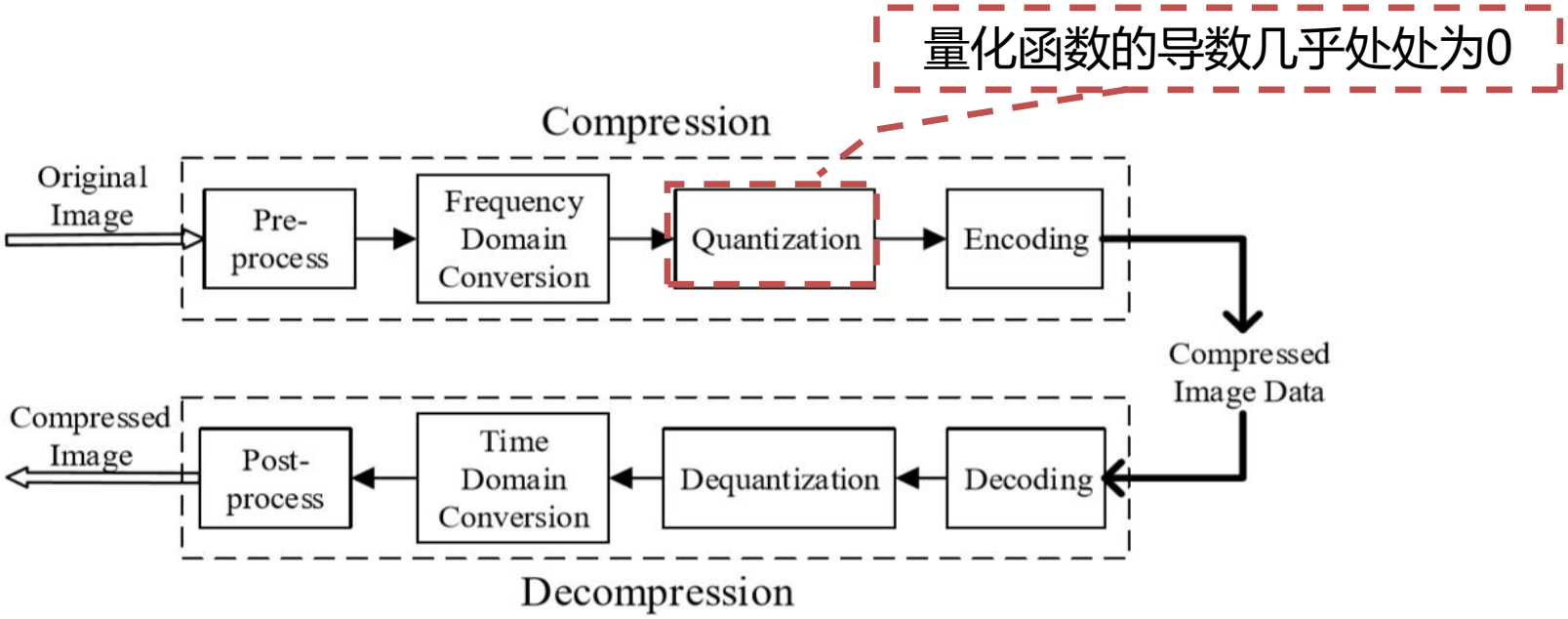


对抗性图像经过有损压缩后易失去对抗性

- 图像压缩技术广泛应用于社交平台来节约通信资源和提高访问效率，但其能破坏精心构造的对抗性扰动，导致对抗性图像失效

对抗性图像需要抗压缩，才能应用于真实社交网络保护图像隐私

图像压缩算法的不可微或未知导致难以生成抗压缩的对抗性图像



有损压缩的主要过程

1 典型的图像压缩方法是**不可微的**，与基于梯度的对抗性图像生成算法不兼容

2 社交平台上的压缩方式是自定义的而不是标准的压缩算法，对于用户**未知**

图像压缩算法的不可微或未知导致难以生成抗压缩的对抗性图像

- **隐私保护**
  - ×针对图像分类模型 [Seong et al. 2017]
  - ×针对目标检测模型 [Liu et al. 2017]
  - ×针对人脸识别模型 [Shawn et al. 2020]

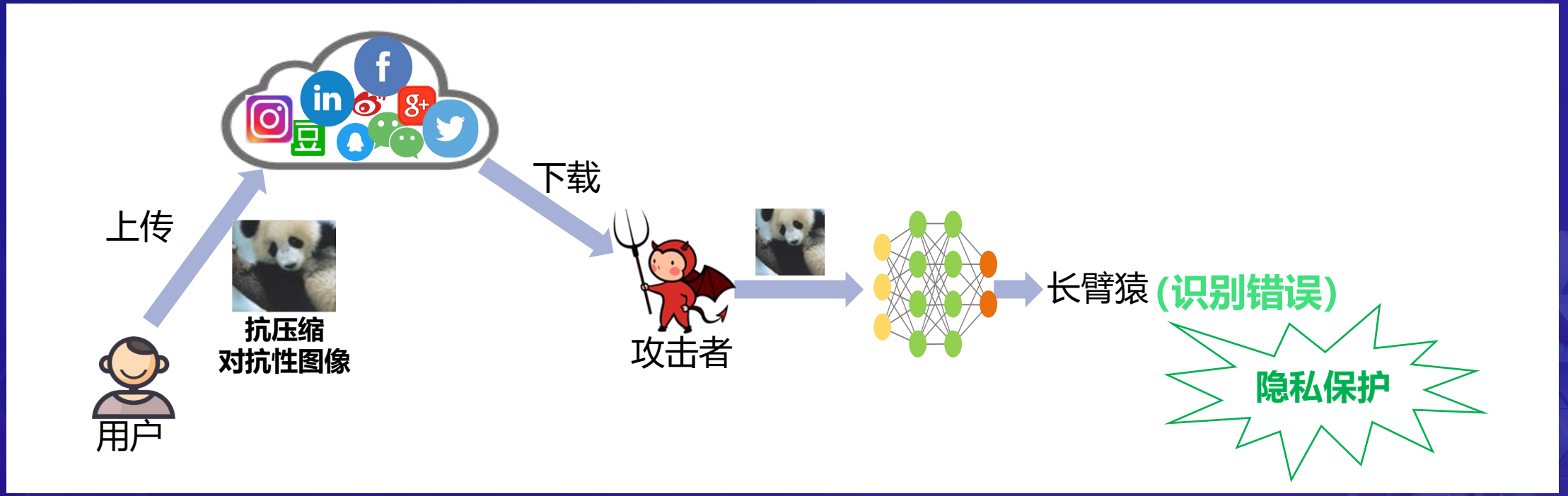
- **抗JPEG压缩对抗性图像** [Richard et al. 2017]
  - ×重写不可微的JPEG压缩算法至可微形式
  - ×要求JPEG压缩算法，相关参数已知
  - ×操作复杂，难以扩展至其他压缩算法
  - ×无法适用于社交平台的未知压缩方式

现有方案无法实现真实社交平台下的图像隐私保护

Seong J, et al. Adversarial Image Perturbation for Privacy Protection A Game Theory Perspective (ICCV'17)
Liu Y, et al. Protecting Privacy in Shared Photos via Adversarial Examples Based Stealth (SCN'17)
Shawn S, et al. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models (USENIX'20)
Richard S, et al. JPEG-resistant adversarial images (NIPS'17)

提出一种抗压缩的对抗性图像生成方案，实现社交网络上的图像隐私保护

- 分享对抗性图像，误导攻击者获得错误的隐私信息
- 对抗性图像具备抗压缩性，上传下载后依旧保持其对抗性
- 适用于不同的社交平台，具有抗未知压缩的普适对抗性



提出一种抗压缩的对抗性图像生成方案，实现社交网络上的图像隐私保护

❑社交平台采用的压缩算法未知

- 社交平台上的压缩方式是自定义的，既不是标准的压缩算法，也不公开，对于我们来说是“黑盒”的

❑压缩算法不可微

- 不可微的压缩算法和基于梯度的对抗样本生成算法不兼容

❑抗压缩对抗性图像生成

挑战一：如何获得“黑盒”压缩算法的近似形式

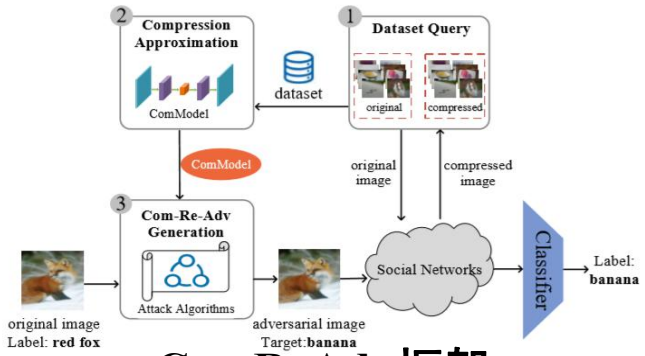
挑战二：如何将不可微的压缩算法转化成可微形式

挑战三：如何有效生成抗压缩的对抗性图像

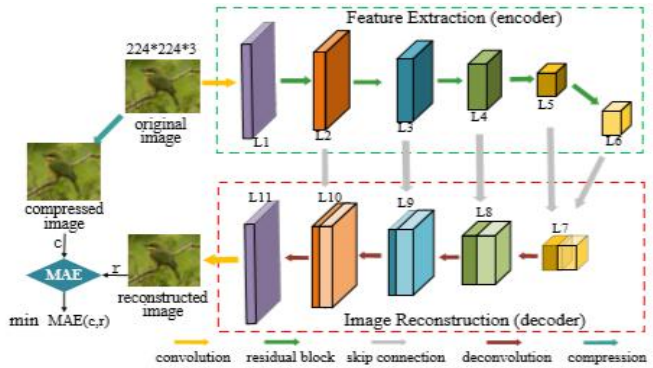
模型假设

- **能直接访问社交网络上的图像**
  - ✓社交平台的共享性质
- **社交平台上的图像压缩方式是“黑盒”的**
  - ×用户对于社交平台采用的压缩方式一无所知（压缩算法及相关参数）
- **图像识别模型是白盒已知的（模型结构和参数）**
- **保持较好的图像视觉效果**
  - ✓保证分享图像可被正常浏览

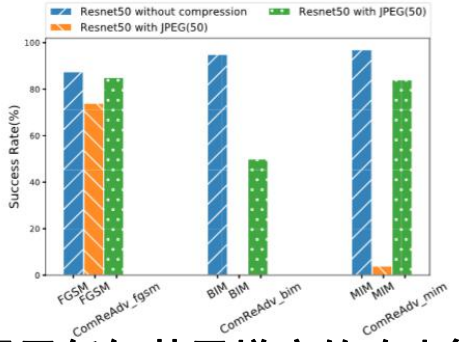
提出了基于编码器-解码器模型近似压缩的抗压缩对抗样本生成技术，首次实现社交平台未知压缩方式下的抗压缩对抗性图像生成

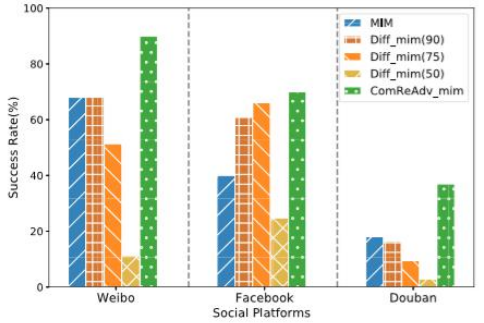


ComReAdv框架

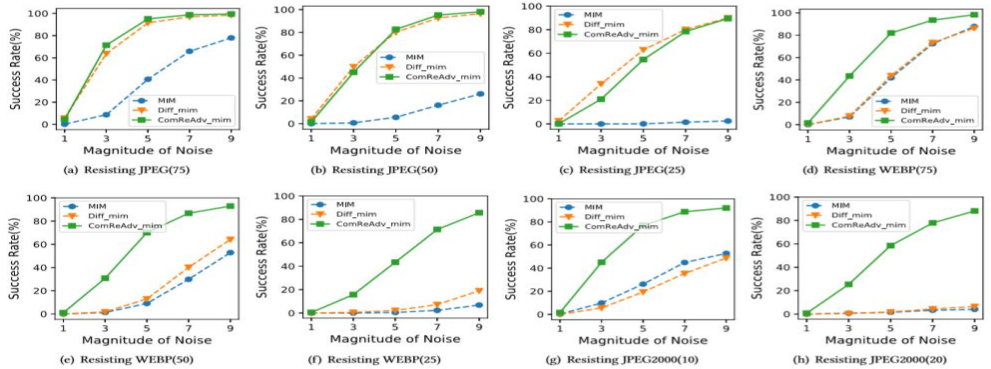


压缩近似模型ComModel





适用于任何基于梯度的攻击算法　现实场景下性能**更优**

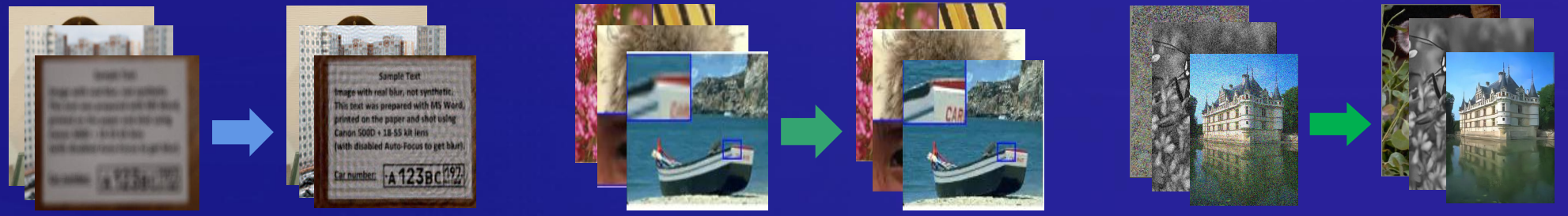


对于任意压缩方式，均取得**高**抗压缩攻击成功率

Zhibo Wang, et al. Towards Compression-Resistant Privacy-preserving Photo Sharing on Social Networks. **MobiHoc 2020**. **(CCF B)**

深度神经网络在图像域迁移上已经取得了优异的效果

图像域迁移



思路

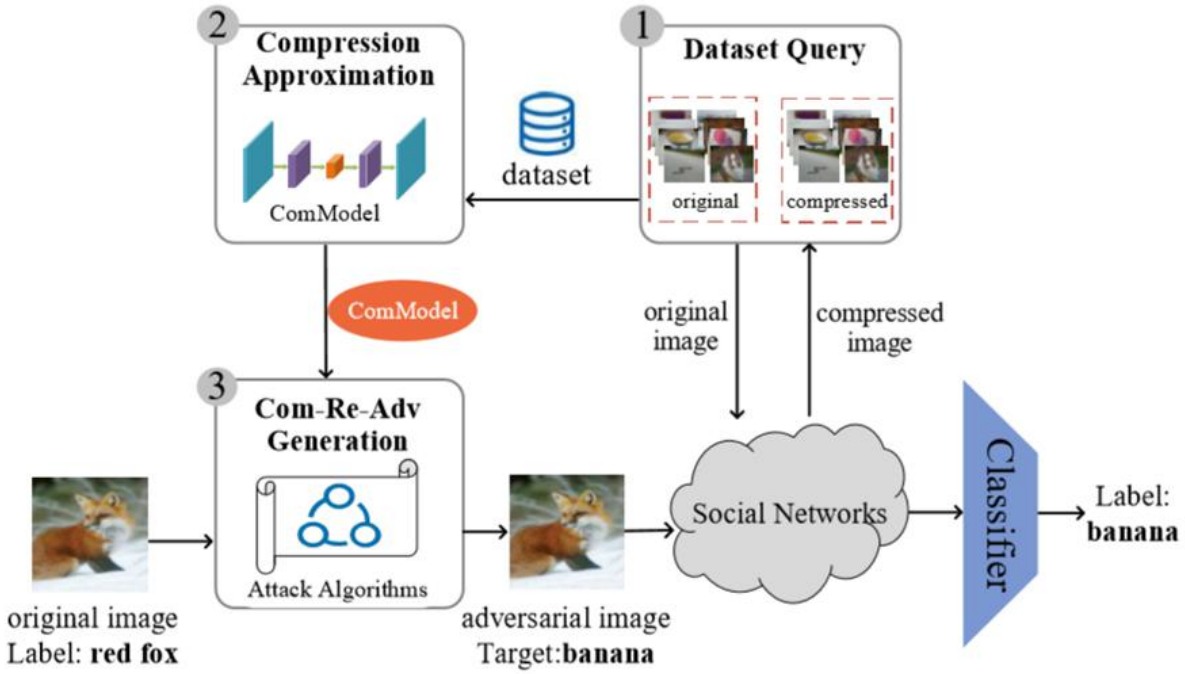
- 将原图到压缩图的转换看成是一种图像域迁移
- 利用深度神经网络学习原图到压缩图的转换
- 将训练好的模型做为压缩算法的近似形式

优点：

- 可微
- 无需了解压缩算法细节
- 端到端学习，容易实现
- 近似效果好

抗压缩对抗性图像生成

- 设计了基于编码器-解码器的压缩近似模型，利用原始图像-压缩图像数据集训练学习压缩算法的近似形式
- 基于压缩近似模型改进现有对抗样本生成算法，生成抗压缩对抗性图像



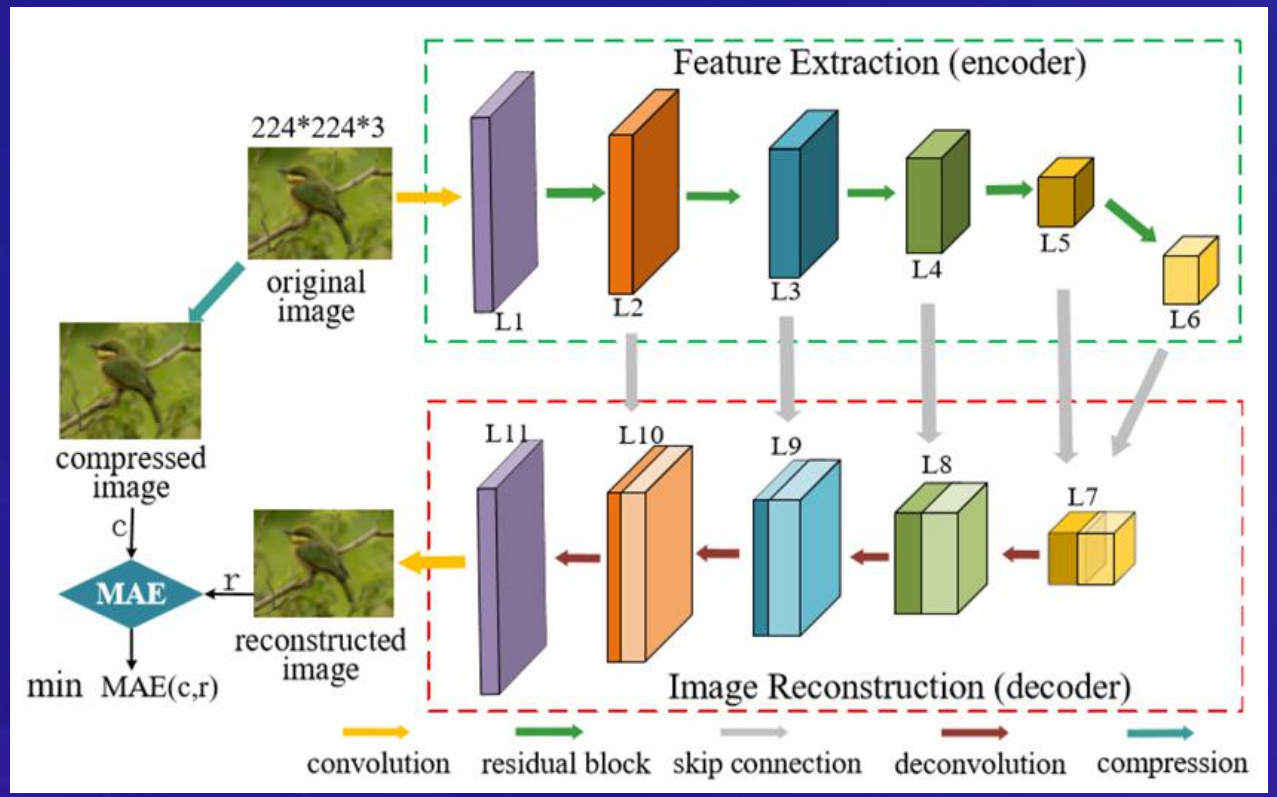
① **原始-压缩数据集构建：**通过上传/下载获取原始图像和压缩图像的数据集

② **压缩算法近似：**训练设计的ComModel模型学习如何转换原图到压缩图

③ **抗压缩对抗性图像生成：**改进现有算法，生成抗压缩的对抗性图像

压缩近似模型

- **目标**：学习原图到压缩图的转换，达到像素级相似，以便作为压缩算法的近似形式

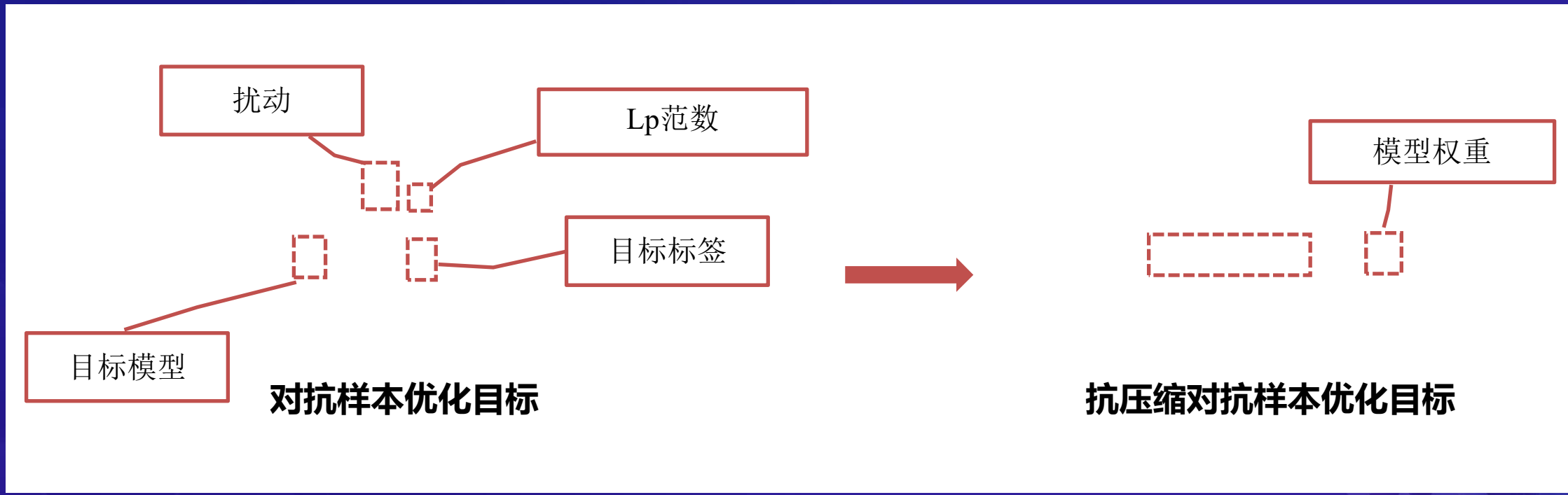


- **编码器**：提取丰富的高层语义特征
- **解码器**：重建图像，使其与真实压缩图在像素级上相似
- **带步长的(反)卷积**：自适应降维，增强输出图像的细节信息
- **残差模块**：避免梯度消失，提高训练效率
- **跳跃连接**：增强输出图像的纹理效果

抗压缩对抗性图像生成

目标：增强对抗样本的抗压缩性

- 非定向对抗样本：误导模型输出错误的结果
- 定向对抗样本：误导模型输出指定的结果



目标：增强对抗样本的抗压缩性

■ **方法：** 基于FGSM[1]，BIM[2]，MIM[3]的求解思想，求解抗压缩对抗样本优化目标

$$\arg\min_r ||r||_p$$

$$s.t.\ C(ComModel(x';\theta)) = t$$

$$x' = x + r \in [0,1]^n$$

抗压缩对抗样本优化目标

1

ComReAdv_fgsm:

$$r = sign\left(\nabla_x \mathcal{L}\left(C\left(ComModel(x;\theta)\right),t\right)\right),$$

$$x' = clip(x - \epsilon \cdot r)$$

2

ComReAdv_bim:

$$r_{k+1} = sign\left(\nabla_{x_k} \mathcal{L}\left(C\left(ComModel(x_k;\theta)\right),t\right)\right),$$

$$x_{k+1} = clip(x_k - \alpha \cdot r_{k+1}), x_0 = x$$

3

ComReAdv_mim:

$$r_{k+1} = sign\left(m \cdot r_k + \nabla_{x_k} \mathcal{L}\left(C\left(ComModel(x_k;\theta)\right),t\right)\right),$$

$$x_{k+1} = clip(x_k - \alpha \cdot r_{k+1}), x_0 = x, r_0 = 0$$

[1] Ian J Goodfellow, et al. Explaining and harnessing adversarial examples (arXiv'14).
[2] Alexey Kurakin, et al. Adversarial examples in the physical world (arXiv'16).
[3] Yinpeng Dong, et al. Boosting adversarial attacks with momentum (CVPR'18).

训练设置

- 数据集：ILSVRC-2012 Validation Set (共50000张图像)，其中5000张用于训练，1000张用于测试， 1000张用于验证
- 优化器：Adam [1]
- 其他设置: batch_size=32, epochs=100

对比算法

- FGSM [2]
- BIM [3]
- MIM [4]
- JPEG-Resistant Attack [5]

目标模型

- ResNet50

测试场景

- 本地仿真测试：JPEG，WEBP， JPEG2000
- 真实社交平台测试：Facebook， 微博，豆瓣

[1] Diederik P Kingma, et al. Adam: A method for stochastic optimization. (arXiv'2014).
[2] Ian J Goodfellow, et al. Explaining and harnessing adversarial examples(arXiv'14).
[3] Alexey Kurakin, et al. Adversarial examples in the physical world(arXiv'16).
[4] Yinpeng Dong, et al. Boosting adversarial attacks with momentum(CVPR'18).
[5] Richard Shin, et al. JPEG-resistant adversarial images(NIPS'17).

❑模型近似压缩效果

压缩质量：对于JPEG和WEBP，数字越小，压缩程度越大，JPEG2000与之相反

Compression	JPEG(75)	JPEG(50)	JPEG(25)	WEBP(75)	WEBP(50)	WEBP(25)	JPEG2000(10)	JPEG2000(20)
Loss Value(MAE)	1.66	2.04	2.47	2.55	2.88	3.44	2.67	3.17

损失误差：ComModel模型输出与真正压缩图的像素级差异

提出的ComModel模型能有效近似各种图像压缩算法

❑抗压缩方案与原始方案对比

	without compression	with JPEG(50)	
	Original	Original	Proposed (ComReAdv_{*})
BIM	95%	0.5%	50%
MIM	97%	4%	84%

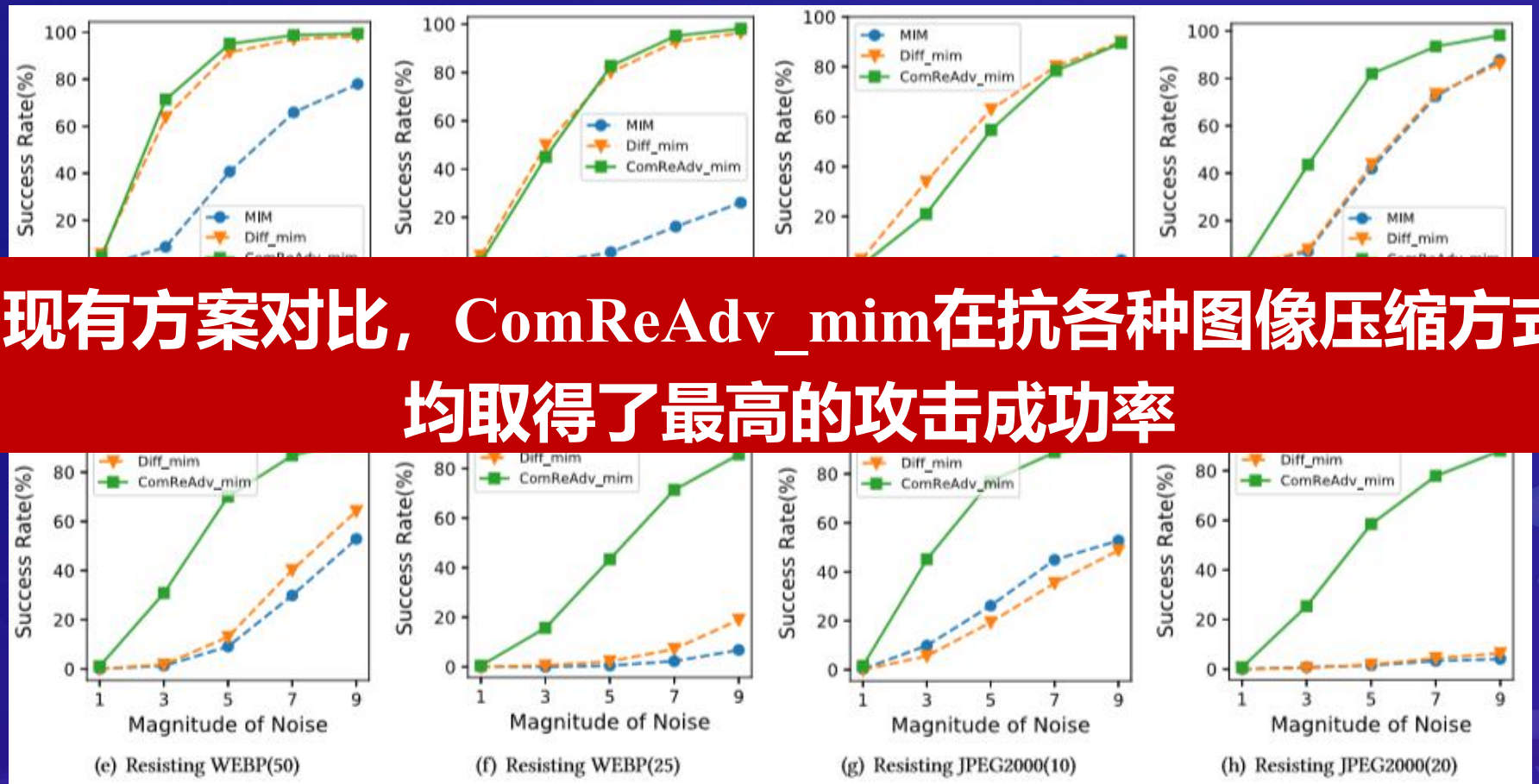
（噪声幅度ϵ =5, 迭代次数10）

ComReAdv_mim
有更好的攻击效果

提出的抗压缩方案能显著提升原始方案的抗压缩效果

❑本地仿真测试

MIM: 原始方案; **Diff_mim**: 抗JPEG压缩方案; **ComReAdv_mim**: 本方案



与现有方案对比，ComReAdv_mim在抗各种图像压缩方式上均取得了最高的攻击成功率

❑现实社交平台测试

现有抗压缩方案

普通方案 | 我们的方案

社交平台 \ 攻击方法	MIM	Diff_mim(90)	Diff_mim(75)	Diff_mim(50)	ComReAdv_mim
微博	68%	68%	51.4%	11.2%	90%
Facebook	40%	60.8%	66%	24.8%	70%
豆瓣	18%	16.4%	9.4%	2.8%	37%

（噪声幅度 $\epsilon = 3$ ，迭代次数10）

我们的方案在现实社交平台上取得了最高的抗压缩成功率

❑结果展示

原始图像

抗压缩对抗性图像

- 微博
- Facebook
- 豆瓣

生成的抗压缩对抗性图像保持原始图像的视觉效果

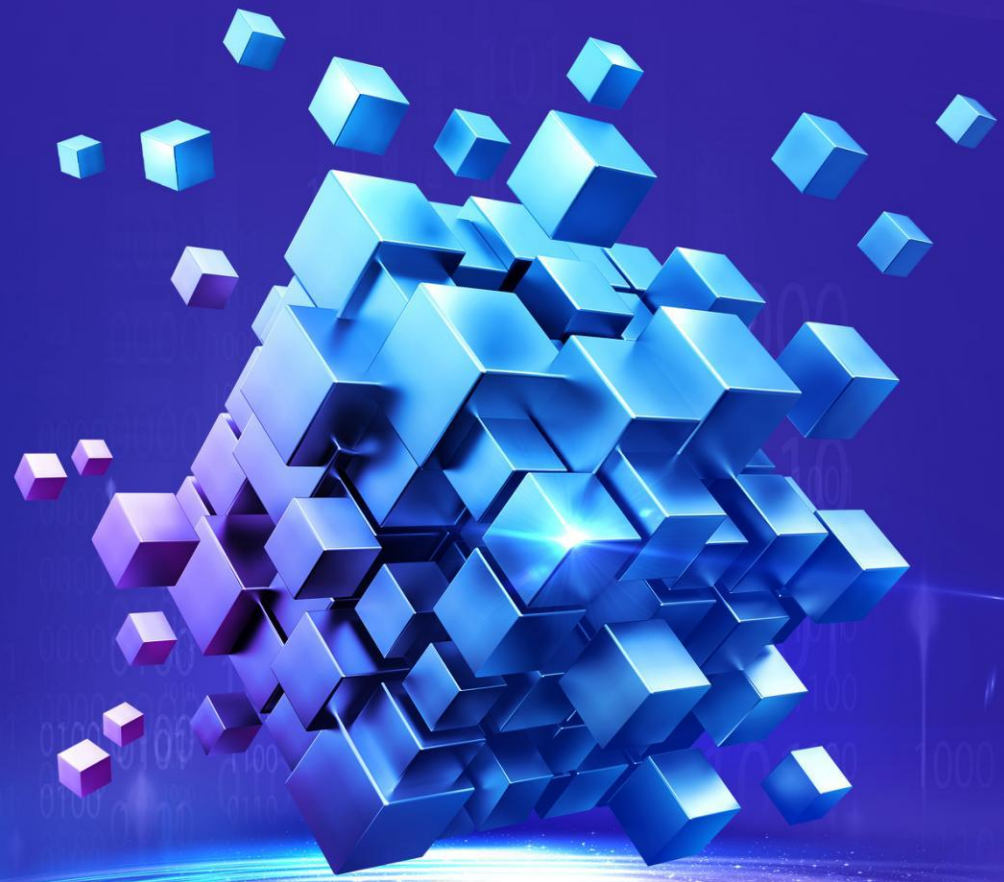
- 设计了适用于任意压缩的抗压缩对抗性图像生成框架
- 提出的框架能结合各种基于梯度的对抗样本生成方法，并显著提高其抗压缩性
- 设计了基于编码器-解码器的压缩近似模型，能有效近似不同压缩算法
- 首次探究了社交网络的压缩对于隐私保护的影响，实现了抗压缩的图像隐私保护
- 对抗性图像不引入明显噪声即可达到误导效果，具有较好的实用性

2021 WEST LAKE CYBERSECURITY CONFERENCE
西湖论剑·网络安全大会

2021 CYBERSECURITY : THE FOUNDATION OF DIGITAL REFORM

敬请批评指正！

西湖论剑·网络安全大会



扫码了解更多西湖论剑资讯