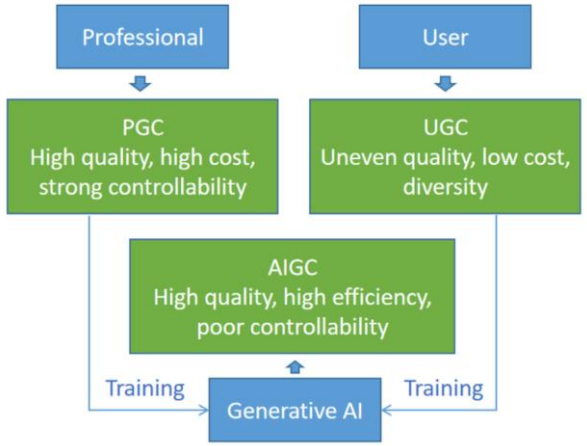

Security and Privacy on
Generative Data in AIGC

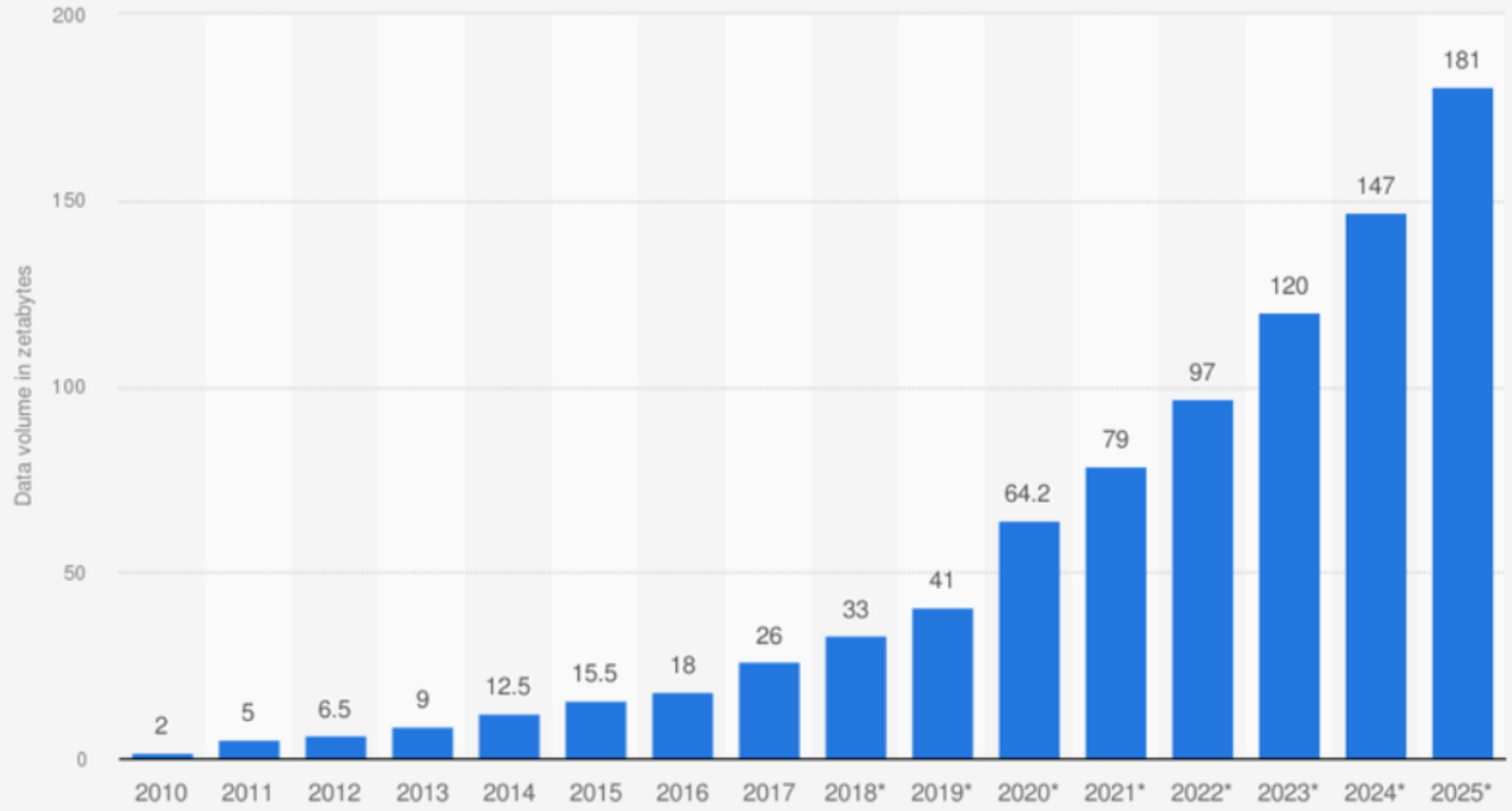
- Three data generation methods, PGC and UGC promote AIGC by providing training data.

A COMPARISON OF CONTENT CREATION MODES

	Content Creation Mode	Features	Limitations	Representatives
Web 1.0	PGC	High quality Low diversity	Limited by capacity of production	Web portal
Web 2.0	UGC	High diversity Medium cost	Limited by content quality	Facebook TikTok
Web 3.0	AIGC	High efficiency Near zero marginal cost	Limited by AIGC technology maturity	Metaverse



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (in zettabytes)



Sources
IDC; Seagate; Statista estimates

Additional Information:
Worldwide; 2010 to 2020

Countermeasures

- Memorization rejection in the training loss
- Deduplicating training datasets
- Differential privacy
  - provable, retraining, reducing data utility
- Detecting replicated content
- Machine unlearning
  - Forget-me-not

复制训练数据来暴露个人隐私内容。

生成虚假内容

- for privacy
  - Face Privacy
  - Beyond Face Privacy
- 恶意目的生成数据
  - 侵权：IP问题
    - C：加扰动控制访问(Glaze)
  - 有毒
    - 暴力色情，公平，伦理，偏见，歧视，道德，政治化
      - C:数据集过滤，生成指导，模型微调(concept消融）过滤生成结果
  - 虚假(幻觉)
    - 医疗教育新闻误导（要符合事实常理）
      新型犯罪活动的出现，如人工智能欺诈、诽谤、身份盗窃和冒充

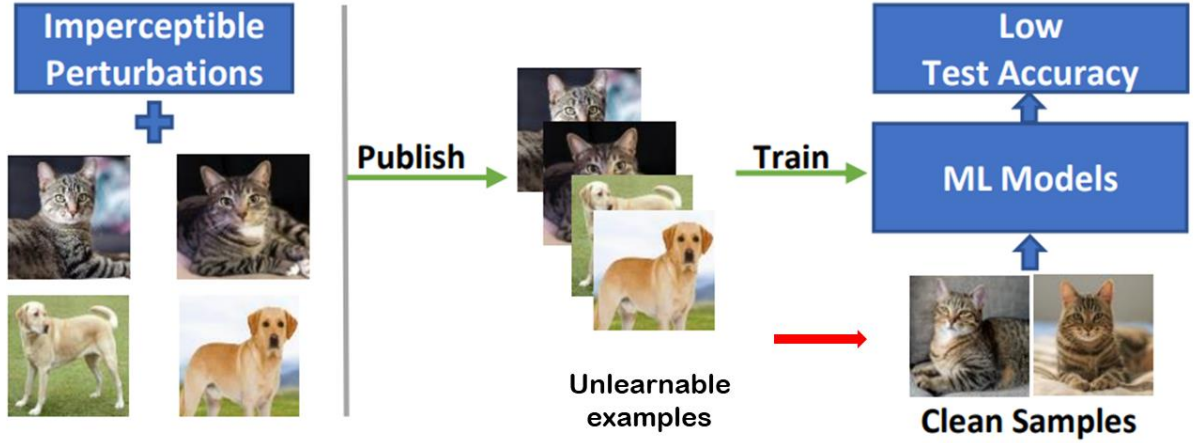
方法：

1. 复制检测和重复数据删除差分隐私
2. Machine Unlearning
3. 控制访问==对训练数据添加扰动（生成数据由训练数据间接引导）
4. 追溯性：水印，区块链
5. 被动保护：生成检测deepfake，生成归因（识别源模型）

Unlearnable Examples Give a False Sense of Security: Piercing through Unexploitable Data with Learnable Examples ACM MM 2023

背景介绍

- 网络上充斥着大量可自由访问的数据，这些数据可能携带未经授权收集的个人信息，引发了公众对隐私的担忧。
- 为了解决这些难题，越来越多的研究力量正在集中于使数据无法被滥用的方向。
- 比如向图像中引入难以察觉的“捷径”噪声，在这种数据上的训练得到的模型，无法准确分类干净的数据，有效地保护了用户的隐私。这种巧妙的方法被称为**不可学习样本（UE），**也可称之为可用性攻击。



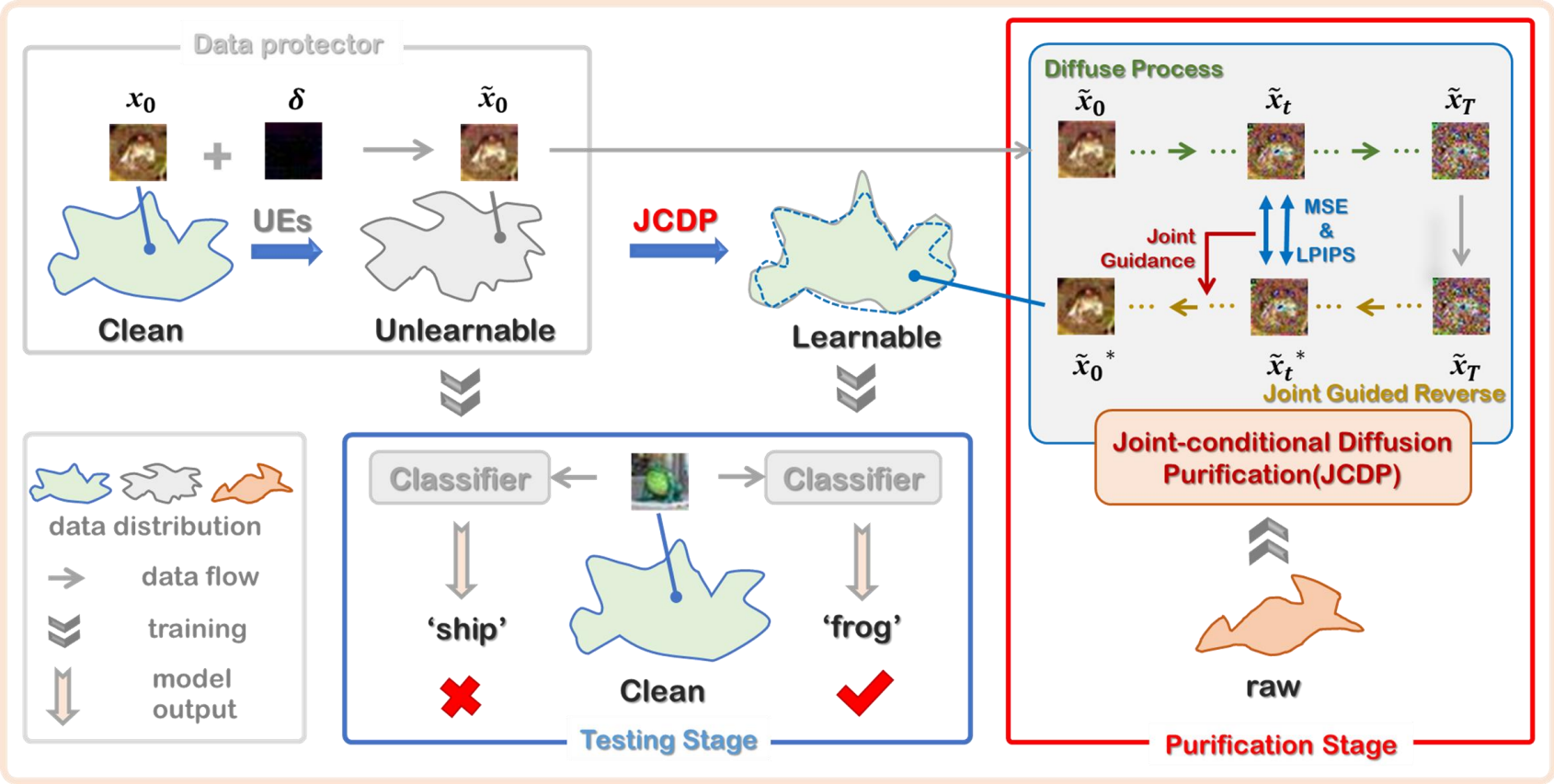
背景介绍

- 我们发现了在这种保护中的关键漏洞：
  - 数据保护人员只能在他们自己的数据中添加“不可学习”的扰动，却无法阻止未经授权的用户访问其他来源的类似的未受保护数据。
  - 未经授权的用户可以很容易地绕过数据保护，从新收集的未受保护数据中学习原始数据表示，即使这些数据可能规模很小，与干净的数据不同，缺乏标签注释，并且单独不适合训练分类器。

- 为了证明上述漏洞的存在，我们设计了一种新的方法，可以将不可学习的样本转化为可学习的样本。

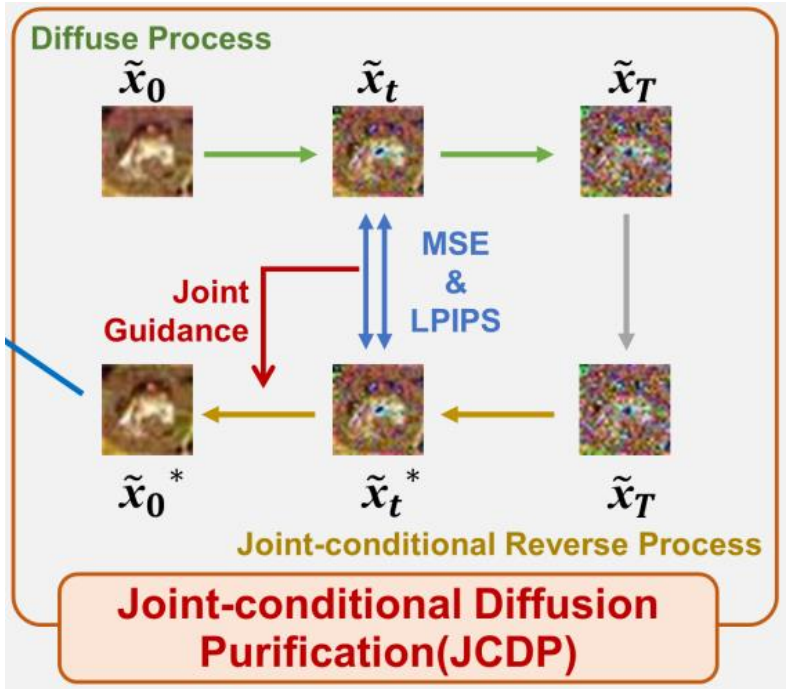
挑战

- 训练数据不可用，需要收集数据
- 原始训练数据和新收集数据的分布不一致性
- 净化和语义保留之间的平衡



- 从其他类似数据中学习一个可学习的数据流形，然后将不可学习的示例投射到该流形上。
- 提出了一种新的联合条件扩散净化方法，以捕获从不可学习样本到相应的干净样本的映射。

JCDP



$$p(\tilde{x}_t|\tilde{x}_t^*) = \frac{1}{Z}\exp(\lambda_1 \mathcal{D}_m(\tilde{x}_t^*, \tilde{x}_t))$$
$$\mathcal{D}_m(\tilde{x}_t, \tilde{x}_t) = \|\tilde{x}_t^* - \tilde{x}_t\|_2$$
MSE

$$p(\Phi(\tilde{x}_t), |\tilde{x}_t, \tilde{x}_t^*) = \frac{1}{Z}\exp(\lambda_2 \mathcal{D}_p(\tilde{x}_t^*, \tilde{x}_t))$$
$$\mathcal{D}_p(\tilde{x}_t, \tilde{x}_t) = \|\Phi(\tilde{x}_t^*) - \Phi(\tilde{x}_t)\|_2$$
LPIPS

- **The diffusion process:**

$$q(\tilde{x}_{1:T}|\tilde{x}_0) = \prod_{t=1}^{T} q(\tilde{x}_t|\,\tilde{x}_{t-1}) \qquad T\text{: diffusion step}$$

- **The Joint-conditional reverse process:**

given a DDPM $(\mu_\varphi, \sigma_t^2 \mathrm{I})$

the conditional transition operator

$$p_\varphi(\tilde{x}_{t-1}^*|\tilde{x}_t^*, \tilde{x}_t, \Phi(\tilde{x}_t)) \approx \mathcal{N}(\mu_\varphi + \sigma_t^2(\boldsymbol{d_1} + \boldsymbol{d_2}), \sigma_t^2 \mathrm{I})$$

obtain the denoised learnable sample

$$\tilde{x}_{t-1}^* \sim \mathcal{N}(\mu_\varphi + \sigma_t^2(\boldsymbol{d_1} + \boldsymbol{d_2}), \sigma_t^2 \mathrm{I})$$

$$\boldsymbol{d_1} = \nabla_{\tilde{x}_t^*} \log p(\tilde{x}_t|\tilde{x}_t^*) \qquad \boldsymbol{d_2} = \nabla_{\tilde{x}_t^*} \log p(\Phi(\tilde{x}_t), |\tilde{x}_t, \tilde{x}_t^*)$$

$$\boldsymbol{d_1} = -\lambda_1 \nabla_{\tilde{x}_t^*} \mathcal{D}_m(\tilde{x}_t^*, \tilde{x}_t) \qquad \boldsymbol{d_2} = -\lambda_2 \nabla_{\tilde{x}_t^*} \mathcal{D}_p(\tilde{x}_t^*, \tilde{x}_t)$$

实验结果

Evaluation on Supervised UEs

Countermeasures	CIFAR-10 (Clean 95.3)				CIFAR-100 (Clean 78.8)				SVHN (Clean 96.2)		
	EMN [22]	EMN (C) [22]	REMN [12]	LSP [48]	EMN [22]	EMN (C) [22]	REMN [12]	LSP [48]	EMN [22]	REMN [12]	LSP [48]
Vanilla	21.2	20.7	20.5	15.0	14.8	4.0	10.9	4.1	13.9	-	7.3
AVATAR [7]	91.0	-	88.5	85.7	65.7	-	64.9	58.5	93.8	88.5	83.8
ISS [26]	93.0	-	**92.8**	82.5	67.5	-	57.3	53.5	89.9	-	92.2
AT [27]	84.8	85.0	49.2	80.2	63.4	60.1	27.1	58.1	86.3	70.0	80.2
AA [34]	90.8	-	85.5	84.9	70.0	-	-	67.4	88.7	-	92.6
LE (Ours)	**93.1**	**94.0**	92.2	**92.4**	**70.9**	**67.8**	**65.3**	**68.7**	**94.7**	**89.9**	**93.3**

Evaluation on Unsupervised UEs

Data	Backbone	Clean	CP [17]	
			Vanila	LE (Ours)
CIFAR-10	SimCLR	90.4	44.9	**86.6**
	MoCo v2	89.3	55.1	**86.0**
	BYOL	92.2	59.6	**85.7**
CIFAR-100	SimCLR	63.6	34.7	**57.4**
	MoCo v2	65.2	41.9	**57.1**
	BYOL	65.3	39.2	**57.2**

Performance Analysis

- 分布相似性

Setting	Data	S-Distribution	EMN	LSP
(1)	CIFAR-10	Vanilla	21.2	15.0
		CIFAR-100	92.1	89.1
(2)	CIFAR-100	Vanilla	14.8	4.1
		CIFAR-10	66.9	66.0
(3)	CIFAR-10	SVHN	89.3	85.6
	CIFAR-100	SVHN	52.9	54.1

Note: LE 可以在很大程度上容忍分布不匹配

- **Countering against Stronger UE protection**

Scale	Standard	AA [34]	AT [27]	ISS [26]	LE(Ours)
8 /255	21.2	90.8	86.2	93.0	**93.1**
16/255	22.6	86.7	83.1	63.4	**87.3**
24/255	21.1	79.3	82.4	-	**83.9**

- **Model Transferability**

Model	Clean	EMN			LSP		
		Vanilla	AA	LE	Vanilla	AA	LE
Resnet-50	94.4	25.2	89.6	**93.2**	14.9	84.2	**92.7**
DenseNet-121	95.1	34.9	91.2	**93.1**	22.7	86.2	**92.3**

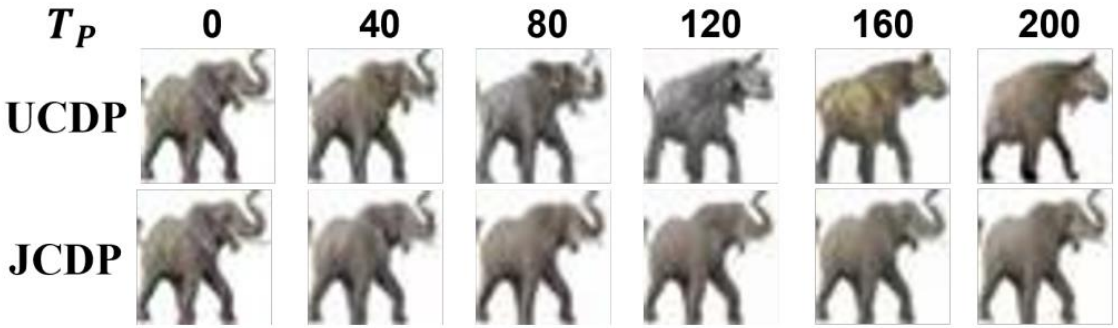
实验结果

消融实验

- **Fine-tuning vs. Training from Scratch**

FT	JC	Steps	EMN(CIFAR-10)	EMN(CIFAR-100)
✗	✗	80000	90.6	69.0
✓	✗	1000	91.4	69.3
✓	✗	10000	91.9	69.7
✓	✓	10000	**93.1**	**70.9**

- **Joint-conditional Diffusion Purification vs. Unconditional Diffusion Purification**



Motivation：AI生成的滥用容易生成不良照片

擦除工作：

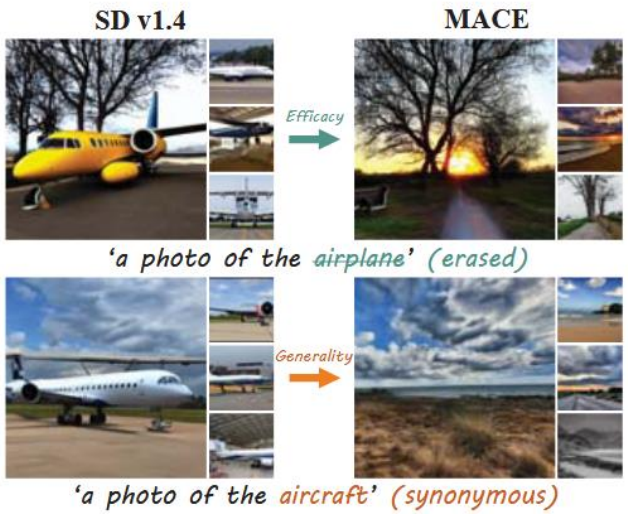
Methods	Accepted	Institution[1]	Contribution
Feature Unlearning for Pre-trained GANs and VAEs	AAAI2024	CSE, POSTECH	识别与目标特征对应的潜在表示，然后使用该表示对预训练模型进行微调。
RECE	ECCV 2024	Fudan University	将不良文本的投影矩阵与无害文本的对齐，再使用迭代生成更好的擦除图片。
MACE	CVPR2024	Nanyang Technological University, Singapore	微调目标短语残余信息的投影矩阵，并使用概念局部重点采样以减少使用多个LoRA模块融合。

2. mass

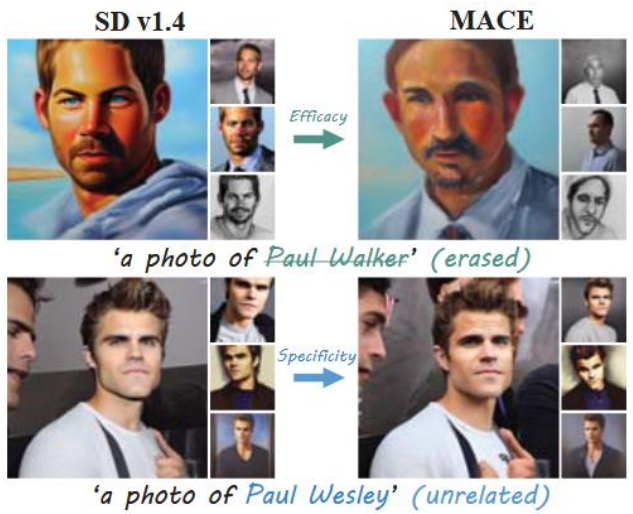
CVPR2024

MACE: Mass Concept Erasure in Diffusion Models

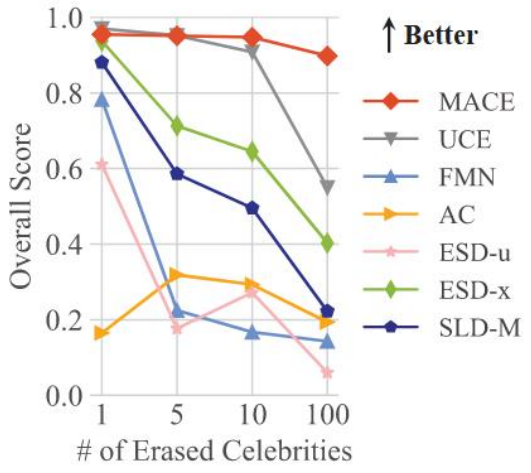
Shilin Lu[1] Zilan Wang[1] Leyang Li[1] Yanzhu Liu[2] Adams Wai-Kin Kong[1]
[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Institute for Infocomm Research (I²R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
{shilin002, wang1982, lile0005}@e.ntu.edu.sg, liu_yanzhu@i2r.a-star.edu.sg, adamskong@ntu.edu.sg



(a) Object Erasure



(b) Celebrity Erasure



(c) Scaling Erasure up to 100 Concepts

Feature Unlearning for Pre-trained GANs and VAEs（AAAI2024）

核心思想：与常见的忘记任务不同，忘记目标是训练集的一个子集，我们的目标是从预训练的生成模型中忘记特定的特征，例如面部图像中的发型。为了指定哪些特征要忘记，我们收集包含目标特征的随机生成的图像。然后，我们识别与目标特征对应的潜在表示，然后使用该表示对预训练模型进行微调。

遗忘框架:

1. 从生成的图像中收集正数据集和负数据集。
2. 在潜空间中找到一个表示目标特征的潜在表示z_e。
(计算正数据集和负数据集的均值向量，并做差，进而得到的目标向量 z_e 用于在潜空间中表示目标特征。)
3. 从一个简单分布中抽取一个潜在向量z。
(a)如果潜在向量不包含目标特征，则不进行修改，让生成器产生相同的输出。
(b)如果潜在向量包含目标特征，则微调生成器以产生不含目标特征的转换输出。
4. 重复步骤3，直到生成器没有生成目标特征。

$$\mathrm{sim}(\mathbf{z}, \mathbf{z}_e) = \begin{cases} 0, & \text{if } \mathrm{proj}_{\mathbf{z}_e}(\mathbf{z}) < t, \\ 1, & \text{otherwise,} \end{cases} \quad (1)$$

投影可以表示潜在向量与目标特征之间的相似性。然后，我们将该值与阈值进行比较，以确定图像是否包含目标特征。

$$\mathcal{L}_{\text{recon}}(\theta) = (1 - \mathrm{sim}(\mathbf{z}, \mathbf{z}_e)) \|g_\theta(\mathbf{z}) - f(\mathbf{z})\|_1, \quad (2)$$

重建损失：当潜在向量不包含目标特征时，unlearning模型g_θ试图模仿原始生成器。
其中，g_ θ是要遗忘的模型，f是预训练生成器。

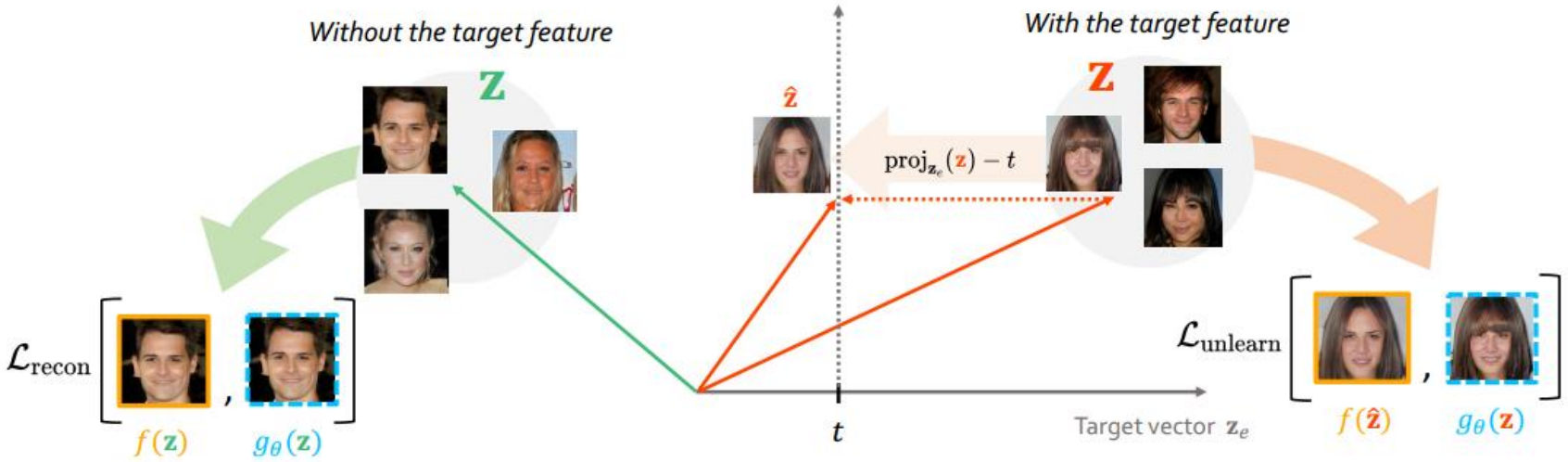
$$\mathcal{L}_{\text{unlearn}}(\theta) = \mathrm{sim}(\mathbf{z}, \mathbf{z}_e) \\ \|g_\theta(\mathbf{z}) - f\left(\mathbf{z} - \left(\mathrm{proj}_{\mathbf{z}_e}(\mathbf{z}) - t\right)\mathbf{z}_e\right)\|_1 . \quad (3)$$

遗忘损失：当潜在向量包含目标特征时，通过调整 z 来改变生成过程，使得g_ θ不再生成包含目标特征的图像。

$$\mathcal{L}_{\text{percep}}(\theta) = \mathrm{sim}(\mathbf{z}, \mathbf{z}_e) \\ \left(1 - \text{MS-SSIM}\left(g_\theta(\mathbf{z}), f\left(\mathbf{z} - \left(\mathrm{proj}_{\mathbf{z}_e}(\mathbf{z}) - t\right)\mathbf{z}_e\right)\right)\right),$$

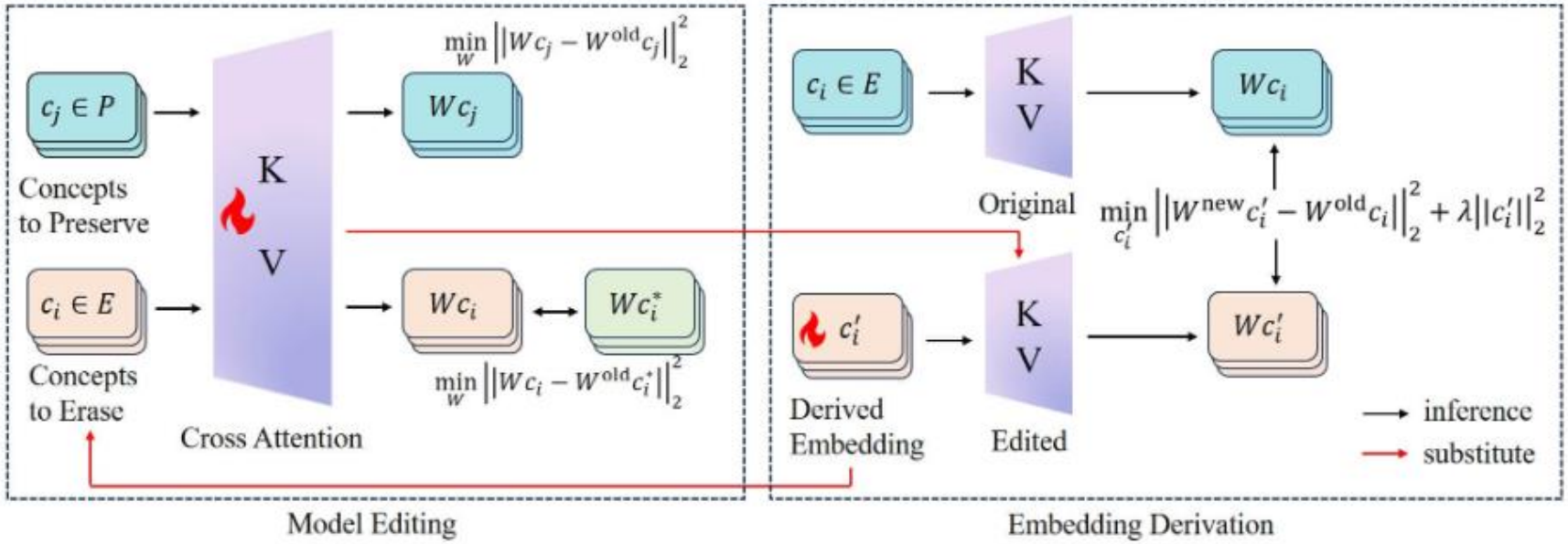
感知损失:当潜在向量包含目标特征时，通过最小化感知损失，可使得模型被鼓励在擦除特定特征的同时，尽可能保持生成图像的视觉质量。

$$\mathcal{L}(\theta) = \alpha\left(\mathcal{L}_{\text{unlearn}}(\theta) + \mathcal{L}_{\text{percep}}(\theta)\right) + \mathcal{L}_{\text{recon}}(\theta), \quad (5)$$



Reliable and Efficient Concept Erasure of Text-to-Image Diffusion Models (ECCV 2024)

RECE高效地利用闭式解来推导新的目标嵌入，这些嵌入能够在未学习的模型内重新生成被消除的概念。为了减轻由推导出的嵌入可能表示的不当内容， RECE进一步将它们与交叉注意力层中的无害概念对齐。新表示嵌入的推导和消除迭代进行，以实现对不当概念的彻底消除。此外，为了保留模型的生成能力，RECE在推导过程中引入了额外的正则化项，从而在消除过程中最小化对不相关概念的影响。



模型编辑和嵌入推导。首先，通过使用闭式解编辑模型来擦除概念，并获得编辑后的交叉注意力W^{new}。然后，给定原始交叉注意力W^{old}和编辑后的W^{new}，通过公式推导出新的嵌入c_i'。在随后的时期，模型编辑和嵌入推导被循环执行。

c_i表示源嵌入（如"裸露"），c_i^*表示相应的目标嵌入（如空文本" "），设E表示要消除的概念，P表示要保留的概念。给定一个K/V投影矩阵W^{old}（W_k^{old}和W_v^{old}的简洁表示），UCE通过编辑E中的概念而保留P中的概念来寻找新权重W。

$$\min_W \sum_{c_i \in E} ||Wc_i - W^{\text{old}}c_i^*||_2^2 + \lambda_1 \sum_{c_j \in P} ||Wc_j - W^{\text{old}}c_j||_2^2 + \lambda_2 ||W - W^{\text{old}}||_F^2, \quad (2)$$

$$W = W^{\text{old}}(\sum_{c_i \in E} c_i^* c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I)(\sum_{c_i \in E} c_i c_i^T + \lambda_1 \sum_{c_j \in P} c_j c_j^T + \lambda_2 I)^{-1}. \quad (3)$$

W^{new}表示UCE编辑后的投影矩阵，c表示"裸露"的嵌入，$c\prime$表示我们导出的嵌入。如果我们能找到一个$c\prime$，使得$W^{new}c\prime$与$W^{old}c$非常相似，那么$c\prime$可以指导编辑后的模型生成裸体图像，就像c指导原始模型一样。

$$\min_{c'} \sum_i \|W_i^{\text{new}}c' - W_i^{\text{old}}c\|_2^2,$$

$$c' = \left(\sum_i W_i^{\text{new}^T} W_i^{\text{new}}\right)^{-1} \left(\sum_i W_i^{\text{new}^T} W_i^{\text{old}}\right) c.$$

复现的效果

使用SDv1.4生成的照片

A portrait of Adam Driver.





使用MACE微调模型进行擦除

A portrait of Adam Driver.



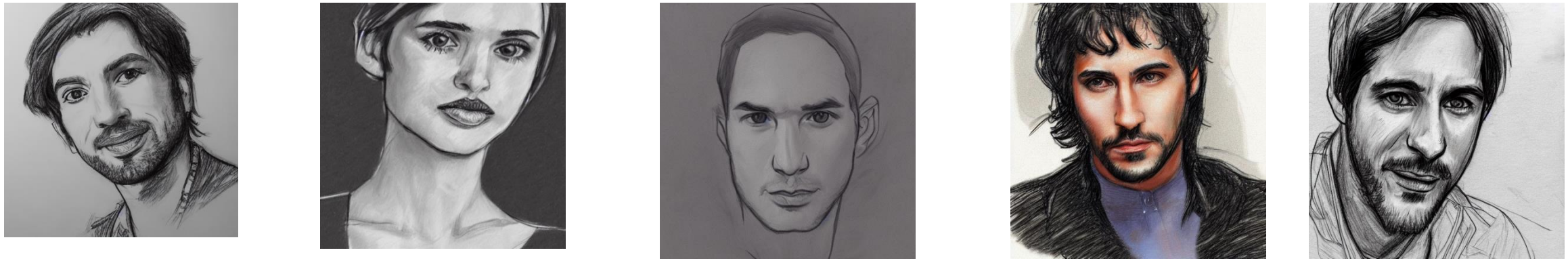


使用SDv1.4生成的照片

A sketch of Adam Driver.

使用MACE微调模型进行擦除

A sketch of Adam Driver.



复现的效果

使用SDv1.4生成的照片

An image capturing Adam Driver at a public event.





使用MACE微调模型进行擦除

An image capturing Adam Driver at a public event.





特异性效果展示

复现的效果

Amy Adams in an official photo.

SDv1.4





mace擦除





Amy Poehler in an official photo.

SDv1.4





mace保留





Security and Privacy on Generative Data in AIGC

Generative Data

情况：
1. 复制训练集数据，学习敏感分布，侵犯隐私
2. 生成的虚假内容，一方面可以替换敏感分布中数据，保护隐私同时保持效用
另一方面存在问题：
侵权，
有毒，（暴力色情，公平，伦理，偏见，歧视，政治化）
虚假，医疗教育新闻误导（要符合事实常理）

控制对其访问：
原因：不受限制访问=>恶意目的（侵权、滥用，产生虚假内容欺骗大众，生成有毒内容（暴力色情伦理偏见歧视））
方法：
1. 控制访问==对训练数据控制（生成数据由训练数据间接引导）
2. 追溯性：水印，区块链
事后被动保护：生成检测，生成归因（识别源模型）