生成式人脸隐私保护赋能安全学习分析

Generative face privacy protection enables secure learning and analysis

王家阳，陈晓雨，刘毓，王涛*
华中师范大学 人工智能教育学部 数字教育湖北省重点实验室 湖北 武汉 430079
*tmac@ccnu.edu.cn

【摘要】 为了有效保护个人隐私，特别是面部识别中的身份和敏感属性信息，本研究旨在开发一种既能保障面部隐私又能维持图像质量和实用性的面部隐私保护方法。本研究提出一种基于扩散模型的面部隐私保护框架，它利用身份编码器和属性编码器对面部进行分解，通过拉普拉斯噪声引入差分隐私来增强保密性，最后由去噪扩散隐式模型 DDIM 作为解码器生成高质量的隐私化面部。实验结果表明，该方法不仅在保护隐私方面表现出色，而且保持了图像质量，并与各种计算机视觉任务兼容，成功地在隐私保护和图像质量及功能性之间找到了平衡点。

【关键词】 面部隐私；扩散模型；差分隐私；身份编码器；属性编码器

Abstract: *To effectively preserve personal privacy, particularly the identity and sensitive attribute information in facial recognition, this study aims to develop a facial privacy-preserving mechanism that not only ensures facial privacy but also maintains image quality and practicality. This research introduces a diffusion model-based facial privacy-preserving framework, which leverages an identity encoder and an attribute encoder to effectively decompose facial representations. Additionally, differential privacy is enhanced by introducing Laplacian noise. Finally, the Denoising Diffusion Implicit Model (DDIM) serves as the decoder to generate privacy-preserving facial images. Experimental results demonstrate that this method not only excels in privacy preservation but also preserves image quality, while being compatible with various computer vision tasks, thereby successfully striking a balance between privacy preservation, image quality, and functionality.*

Keywords: facial privacy, diffusion model, differential privacy, identity encoder, attribute encoder

1.引言

在当今数字化教育蓬勃发展的时代，在线学习等教育场景广泛应用先进的人工智能与计算机视觉技术，以实现更精准、高效的教育分析。例如，通过记录和分析学生的人脸数据，能够获取学生的课堂参与度、情绪状态等关键信息(Kew et al., 2022)，进而为个性化教学提供有力支持。这一技术的应用使得教育工作者能够更深入地了解学生的学习过程，及时调整教学策略，提升教学质量。

然而，这种对学生人脸数据的大规模收集和分析也引发了严峻的隐私问题。当学生的面部照片被在线学习平台记录和处理时，这些数据可能面临被非法利用的风险(Ismail, 2025)。不法分子可能会从中提取学生的身份信息，甚至包括年龄、性别等敏感的人口统计信息，这些信息的泄露不仅侵犯了学生的个人隐私权，还可能引发潜在的歧视问题。比如，基于年龄或性别的偏见可能会影响学生获得公平的教育资源和机会，对学生的身心健康和学习发展造成负面影响。

面对这一挑战，全球各国政府积极制定相关法规，以规范教育领域面部数据的收集、存储、处理及共享流程。例如，欧洲的《通用数据保护条例》（GDPR）(Hiller et al., 2019)和

美国的加利福尼亚州《消费者隐私法案》（CCPA）等，都为个人隐私权提供了法律保障。尽管技术进步为教育带来了诸多便利，但在教育场景中，我们必须高度重视学生面部隐私的保护，探索更为先进的技术手段，确保在保护隐私的同时，不影响教育分析的准确性和有效性。

目前，针对包含面部信息的图像或视频的隐私保护方法，主要分为身份隐藏和属性隐藏两类。身份隐藏技术通常采用模糊化、像素化处理或对面部细节进行修改等方式，降低面部识别技术的识别准确性，从而隐藏学生的身份信息。属性隐藏策略则侧重于隐匿面部敏感属性，如年龄、性别等，以消除潜在的歧视风险。

然而，这些传统方法存在明显的局限性。例如，模糊处理或像素化往往会导致图像质量严重下降，使得基于面部图像的后续教育分析，如表情识别、专注度分析等难以准确进行。此外，仅依靠身份隐藏或属性隐藏，无法充分满足教育场景中对隐私保护和面部图像功能可用性的双重需求。为了解决这些问题，我们结合身份隐藏和属性隐藏的任务，利用具有强大稳定性和图像生成多样性的扩散模型，提出了一种新的面部隐私保护方法。该方法旨在提供一种灵活的面部隐私保护范式，实现可配置、机密且实用的面部隐私保护，确保在保护学生面部隐私的同时，保持图像的视觉逼真度，不影响后续的教育分析工作。

本文主要的研究工作如下。

1) 我们提出了一种基于扩散的灵活面部隐私保护方法，该方法结合了身份隐藏和属性隐藏的任务，同时实现了可配置、机密且高质量的面部隐私保护。

2) 我们引入了身份编码器和属性编码器，以有效分解面部表征，这有助于通过自适应管理隐私关注并同时保护面部身份表征和属性表征，来实现灵活的面部隐私保护。

3) 我们将差分隐私融入其中，通过在身份表征以及控制属性变化的参数中引入拉普拉斯噪声，来保护面部身份和敏感面部属性的隐私，从而增强了我们的方法对潜在隐私攻击的机密性和鲁棒性。

4）我们将保护后的表征由强大的扩散生成模型 DDIM 合成具有匿名身份和属性的面部图像。我们的方法能够在稳健保护面部隐私的同时，生成高质量且自然的面部图像。

2.研究内容

为了克服在多样化的面部隐私保护和高性能要求之间做出妥协的挑战，我们提出了一种利用扩散模型的创新性面部隐私保护方法。具体而言，该方法包括三个阶段，如图 1 所示。阶段一：面部解耦。将输入的面部图像及其潜在空间解耦为身份和属性，然后将它们组合成一个统一的语义表征，通过 DDIM 进行重建恢复原始面部。阶段二：面部净化。根据特定的隐私要求，在身份表征和属性转换参数中注入拉普拉斯噪声，从而隐藏面部隐私。阶段三：面部生成。通过扰动的表征由 DDIM 合成具有匿名身份和属性的面部图像。因此，我们的方法能够在稳健保护面部隐私的同时，生成高质量且自然的面部图像。

我们的方法提供了一种灵活的面部隐私保护范式，同时实现了可配置、机密且实用的面部隐私保护。可配置的面部隐私保护意味着我们可以根据现实生活应用的需求，选择性地保留身份、任何感兴趣的属性或它们的任何组合，同时保持其他属性，包括姿势和背景。机密性意味着该方法生成的增强隐私的面部图像具有高度安全性，能够防止各种面部识别模型的识别。实用性则表明该方法输出的增强隐私的图像保留了其在各种下游计算机视觉应用中的实用性，包括面部检测和表情识别。这种能力使得我们的方法能够在现实场景中有效实施，平衡隐私需求与复杂视觉分析任务的操作要求。

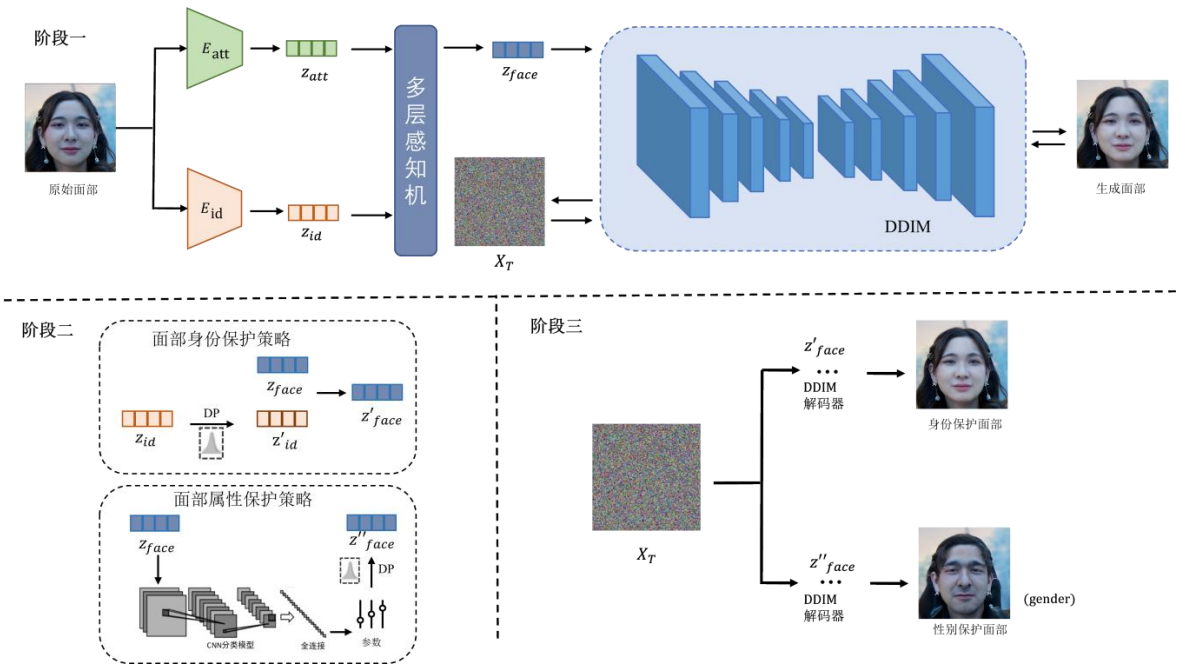

图 1 我们的方法框架概述

3.本文方法设计

3.1. 阶段一：面部解耦

为了有效且灵活地控制多样化的面部隐私，我们开发了两个独立的语义编码器和一个基于扩散的解码器来解耦面部信息。具体而言，身份编码器E_{id}和属性编码器E_{att}分别负责提取与个体身份和面部属性相关的表征z_{id}和z_{att}。这些表征集随后被合并，并通过多层感知机（MLP）进一步处理，以形成一个全面的高级语义表征z_{face}。我们采用了一个条件 DDIM 解码器，它以随机噪声x_T为输入，并使用z_{face}作为条件来指导解码器进行有效的去噪和面部重建 x0 ：

$$p_\theta(x_{0:T}|z_{face} = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z_{face}))$$

为了获得与属性无关的身份表征，我们选择使用预训练的身份识别模型 ArcFace(Deng et al., 2019)作为身份编码器，该编码器将输入图像编码为身份表征$z_{id} = E_{id}(x_0)$。ArcFace 在面部识别领域以其高精度而著称，能够识别不同姿势或表情的面部图像中的个体。因此，我们期望它能提供我们所需的解纠缠特性。多项研究(Nitzan et al., 2020)(Wang et al., 2021)表明，使用预训练编码器提取的表征已经足够解纠缠，无需额外学习。类似地，我们利用一个预训练的属性编码器(Wen et al., 2022),来捕获面部属性信息，如表情、性别和年龄。此外，面部解耦涉及使用此预训练编码器从图像中提取身份和属性表征，然后将这些表征输入到一个额外的可训练 MLP 中，该 MLP 将它们映射到更高的语义空间。这一过程有效地形成了一个全面的面部表征，提高了面部识别和分析的准确性和有效性。

3.2. 阶段二：面部净化

面部隐私保护是一个至关重要的领域，它包含两个关键方面：身份表征保护和敏感属性保护。

身份表征保护：在实践中，我们使用差分隐私技术来加强身份表征的保护。具体而言，我们在身份表征上添加拉普拉斯噪声，引入随机性以有效模糊个体。这种方法确保即使数据被泄露，攻击者也无法准确重建个体的真实身份，从而保护用户的个人安全。

敏感属性保护：为了有效保护如年龄和性别等敏感属性，我们开发了一个高度灵活的属性分类器，能够精确识别这些表征。在我们的研究中，我们采用了 DiffAE(Preechakul et al., 2022)中的方法，并对 CelebA-HQ(Karras, et al., 2017)数据集中的每个属性进行了详细分析。我们将面部图像的身份表征和属性表征相结合，创建了一个高级的语义空间。在这个空间中，我们使用公式$C_{att}(z_{face}) = \text{sigmoid}(w_{att}^{T} z_{face})$训练了一个线性分类器，以根据高级语义表征精确识别特定属性。此外，我们还采用了归一化技术，确保表征保持在单位球面上，具体通过表达式$\lambda_1 \text{Norm}(z_{face} + sw_{att})$来实现。其中，s 是一个缩放超参数，用于控制属性修改的强度。通过调整与特定属性相关的语义表征，并应用缩放和归一化技术，我们旨在在图像中实验性地改变这些属性。

因此，为了保护敏感属性，我们在缩放超参数 s 中引入了拉普拉斯噪声，生成了受扰动的面部语义表征。我们还为属性变化的方向增加了随机性——例如，正值表示年龄增加，而负值表示年龄减少。这种隐私保护策略确保即使在数据公开或共享时，特定的个人信息也能保持机密性。这实现了隐私保护和数据可用性之间的有效平衡。

3.3. 阶段三：面部生成

生成身份匿名化的面部图像：在生成身份匿名化的面部图像的过程中，我们依赖于两个关键元素：已经脱敏的身份表征z'_{id}和原始的属性表征z_{att}。为了确保生成图像的质量，我们在第一阶段训练得到的 DDIM 解码器的所有参数都被固定。通过一个多层感知机（MLP），我们将扰动后的身份表征z'_{id}与原始的属性表征z_{att}相结合，形成匿名化面部的新语义表征z'_{face}。最后，使用这个 DDIM 解码器，我们基于z'_{face}生成高质量的身份匿名化面部图像。这个过程确保了面部身份的保护，同时保持了面部图像的自然性和真实性。

生成属性匿名化的面部图像：生成属性匿名化的面部图像的过程依赖于属性脱敏的语义表征z''_{face}。同样地，为了确保生成高质量的匿名化面部图像，我们在第一阶段训练完成后固定了 DDIM 解码器的所有参数。使用这个具有固定参数的 DDIM 解码器，我们基于语义表征z''_{face}生成属性匿名化的面部图像。

4.实验分析

4.1. 实验设置

数据集：我们在以下数据集上进行保护操作：（i）CelebA-HQ(Karras, et al., 2017)，它包含来自 CelebA 数据集的 30,000 张高分辨率（1024x1024）的名人面部图像，这些图像被标注了 40 个不同的属性标签，涵盖了年龄、性别和种族等人口统计表征。这些标签针对每张图像内外区域的面部表征。（ii）FFHQ 数据集(Karras et al., 2019)，它包含了 70,000 张高质量面部图像，展示了广泛的年龄、性别和表情多样性，所有图像的分辨率均为 128x128 像素。

基准：为了验证本文所提出的方法的有效性，我们将其与基准匿名化方法进行比较：DeepPrivacy(Hukkelås et al., 2019)和 CIAGAN(Maximov et al., 2020)。

攻击模型：为了评估面部身份匿名化的有效性，我们对两个主流的黑盒面部识别（FR）模型进行了全面测试，包括 FaceNet 和 ArcFace。

为了评估面部属性匿名化的有效性，我们专注于保护两个主要属性：性别和年龄。通过一系列实验，我们验证了我们的匿名化算法在这些属性上的有效性。我们采用了两种性别攻击算法：Visual Geometry Group 16 (VGG16)和支持向量机（SVM），以及两种年龄攻击算法（区分年轻和年老）：SVM 和 K 最近邻（KNN）。所有这些算法都是在 CelebA 数据集的原始图像上进行训练的。

评估指标：我们使用攻击成功率（ASR）来评估不同方法的隐私保护效果。此外，我们还采用了几种常见的指标来评估图像质量，包括结构相似性指数（SSIM）、峰值信噪比（PSNR）和学习感知图像块相似性（LPIPS）。

4.2. 面部身份隐私保护

图像质量的结果: 图 2 展示了不同方法生成的匿名面部图像。可以观察到，CIAGAN 生成的匿名图像存在显著的失真，而 DeepPrivacy 虽然保留了一定程度的照片真实性，但未能有效保留与身份无关的属性，如表情。相比之下，我们的方法生成的图像具有自然的面部表征和增强的照片真实性，成功保留了与身份无关的属性，如表情和姿势。表 1 报告了图像质量的评估结果，表明我们的方法在所有图像质量指标上都优于其他方法。

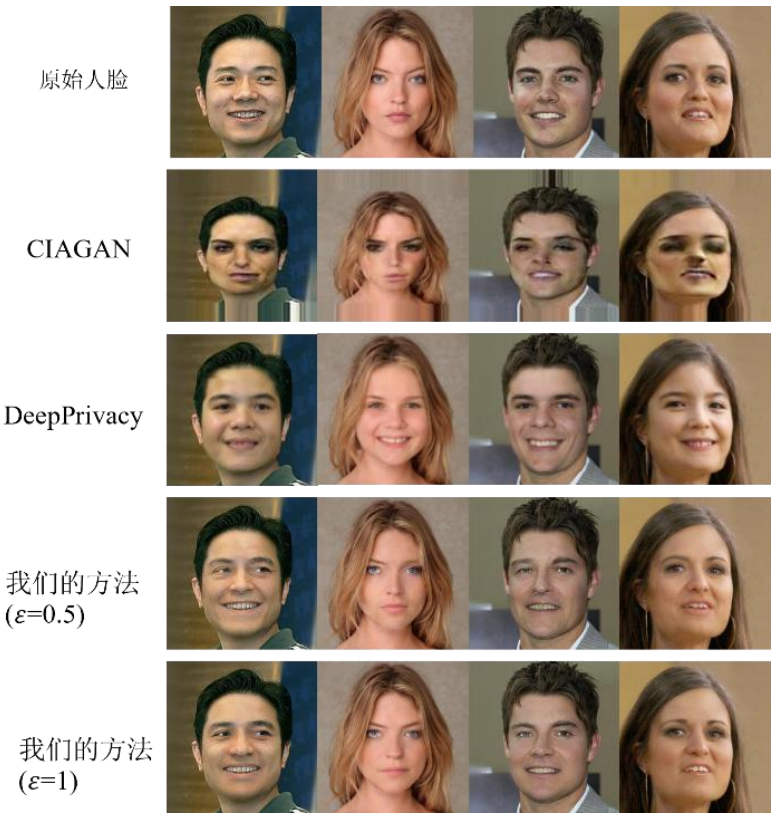

图 2 与基准方法的定性比较

表 1 不同图像匿名化方法图像质量的定量比较

方法	PNSR	SSIM	LPIPS
CIAGAN	21.863	0.7401	0.5499
DeepPrivacy	21.012	0.7808	0.079
我们的方法 ($\epsilon = 0.5$)	27.75	0.8603	0.0330
我们的方法 ($\epsilon = 1.0$)	27.73	0.8635	0.0329

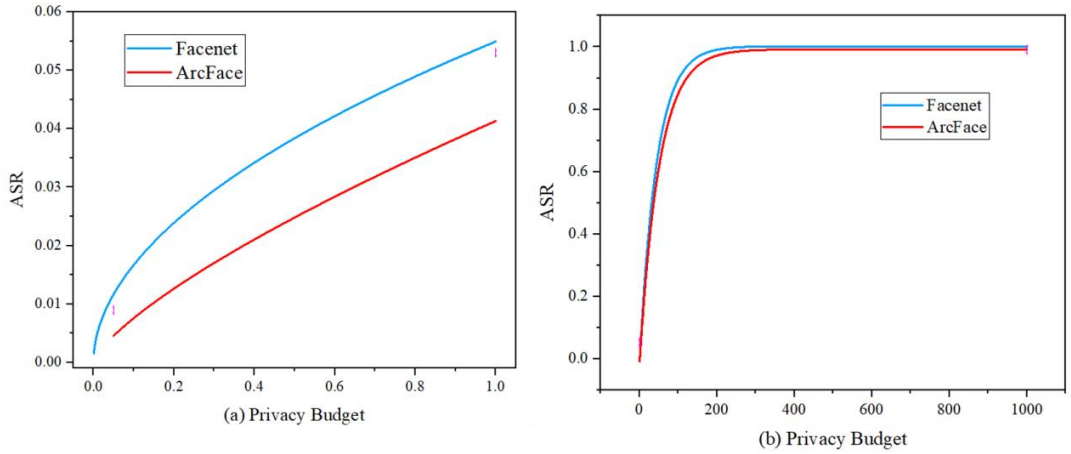

图 3 不同隐私预算下，身份保护图像的黑盒攻击成功率趋势

对黑盒攻击的比较: 表 2 报告了使用 CelebA 数据集对两个流行的面部识别（FR）模型进行黑盒攻击的定量结果。我们测试了针对两个特定身份的目标攻击性能，结果表明，与其他方法相比，我们的方法展示了最低的攻击成功率，这证实了我们的方法在生成匿名面部图像方面的有效性，符合预期。如图 3 所示，随着隐私预算的增加，攻击成功率逐渐上升。这些数据表明，较低的隐私预算在降低成功攻击的可能性方面更为有效，从而增强了匿名化效果。

表 2 不同图像匿名化方法黑盒攻击成功率比较

方法	Facenet 的 ASR	ArcFace 的 ASR
CIAGAN	0.151	0.100
DeepPrivacy	0.077	0.067
我们的方法 ($\epsilon = 0.5$)	0.044	0.024
我们的方法 ($\epsilon = 1.0$)	0.053	0.042

4.3. 面部属性隐私保护

图像质量的结果: 图 4 展示了性别转换的视觉效果，特别是将男性图像转换为女性图像时。生成的女性图像通常具有更柔和的面部表征和更长的头发，这与女性外貌的常见表征相符。图 5 展示了年龄转换的影响：当年轻人被转换为老年人时，生成的老年人图像通常显示出白胡子和更多的皱纹，这是老年人的典型表征。此外，随着隐私预算的增加，生成图像的质量也相应提高。

图 4 性别保护处理后的图像

黑盒攻击的结果：图 6 展示了针对四个流行的黑盒攻击，对属性保护的面部图像（特别是处理过以隐藏敏感属性如性别和年龄的图像）的定量结果。这些结果有效地降低了性别分类器的性能，从而保护了敏感属性。然而，重要的是要注意，在较低的隐私预算下，虽然隐私保护得到了改善，但分类器的性能受到了显著影响。这主要是因为为了增强隐私保护而应用的过多扰动导致图像质量明显下降，进而妨碍了分类器从图像中提取关键面部表征（如眼睛、鼻子和嘴巴的形状）的能力。过多的扰动会模糊或完全遮挡这些关键表征，使得分类器难以提取有意义的信息来进行准确的分类。值得注意的是，这种现象不仅限于性别属性的保护，还适用于其他敏感属性的匿名化，如年龄，这凸显了我们的方法在保护各种类型私人信息方面的多功能性和有效性。

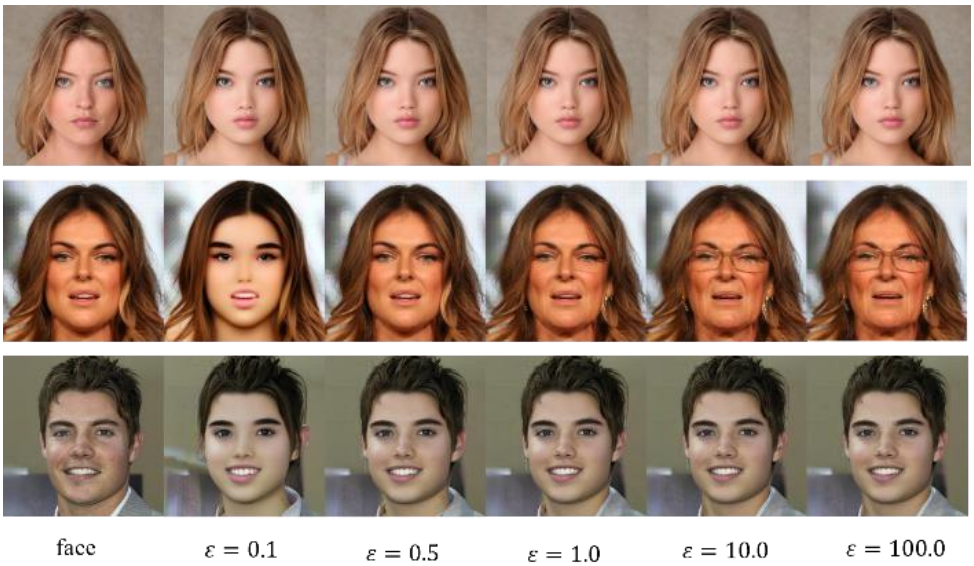

图 5 年龄保护处理后的图像

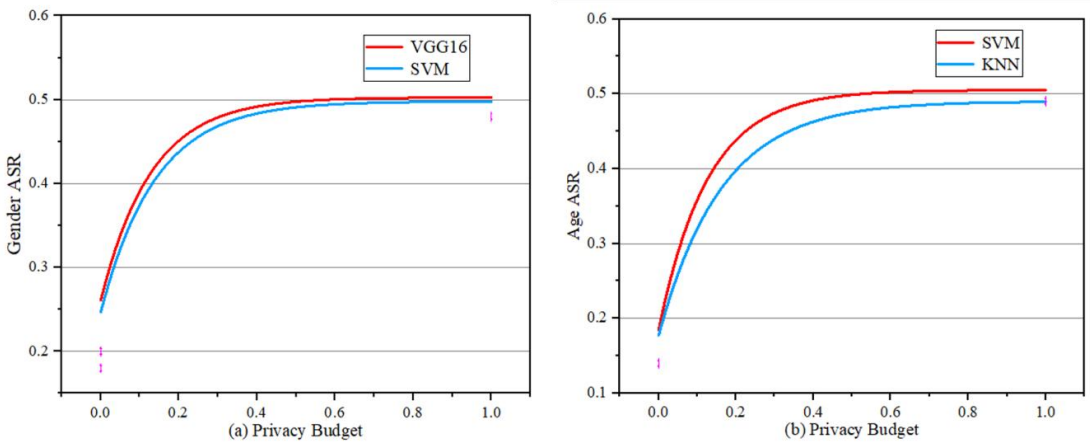

图 6 不同隐私预算下，属性保护图像的黑盒攻击成功率趋势

5.结束语

在本文中，我们提出了一种利用扩散模型的创新性面部隐私保护方法，它巧妙地融合了扩散模型的核心优势，实现了对身份与属性信息的深度隐藏。通过引入身份编码器和属性编码器，结合扩散模型强大的生成能力，我们成功地将面部进行了分解，为隐私保护开辟了新路径。为了进一步增强隐私保护的力度，我们引入了拉普拉斯噪声机制，以此为基础构建了差分隐私框架，显著提升了系统的保密性能。最后，由 DDIM 生成高质量的面部隐私保护图像。实验数据充分证明，实验结果表明，我们的方法不仅在隐私保护方面表现出色，而且保持了图像质量，并与各种计算机视觉任务兼容。

参考文献

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4690-4699).

Hiller, J., Schuldes, M., & Eckstein, L. (2019, October). Recognition and pseudonymization of data privacy relevant areas in videos for compliance with GDPR. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)* (pp. 2387-2393). IEEE.

Hukkelås, H., Mester, R., & Lindseth, F. (2019, October). Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing* (pp. 565-578). Cham: Springer International Publishing.

Ismail, I. A. (2025). Protecting Privacy in AI-Enhanced Education: A Comprehensive Examination of Data Privacy Concerns and Solutions in AI-Based Learning. *Impacts of Generative AI on the Future of Research and Education,* 117-142.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196.*

Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks.* In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

Leslie, D. (2020). Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023.*

Kew, S. N., & Tasir, Z. (2022). Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. T*echnology, Knowledge and Learning, 27*(2), 405-427.

Maximov, M., Elezi, I., & Leal-Taixé, L. (2020*).* Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5447-5456).

Nitzan, Y., Bermano, A., Li, Y., & Cohen-Or, D. (2020). Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728.*

Preechakul, K., Chatthee, N., Wizadwongsa, S., & Suwajanakorn, S. (2022). Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10619-10629).

Wang, T. C., Mallya, A., & Liu, M. Y. (2021). One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10039-10049).

Wen, Y., Liu, B., Ding, M., Xie, R., & Song, L. (2022). Identitydp: Differential private identification protection for face images. *Neurocomputing, 501*, 197-211.