Identity-Preserving Face Anonymization via Adaptively Facial

Attributes Obfuscation

**摘要**

随着计算机视觉技术在监控系统中的应用越来越普遍，由于监控系统所拍摄的图像或者视频中可能包含如人脸等与身份相关的信息，大家越来越关注监控系统所带来的个人隐私侵犯问题。现有保护此类隐私的方法主要侧重于从人脸图像中去除与身份相关的信息，但是这样会削弱当前监控系统的有效性。在本文中，我们开发了一个人脸匿名化框架，该框架在混淆视觉外观的同时能够保持身份的可辨别性。此框架由两部分组成：身份感知区域发现模块和身份感知的人脸混淆模块。前者自适应地定位人脸中与身份信息无关的属性，后者利用原始人脸和发现的身份信息无关属性生成隐私保护人脸。为了优化人脸生成器，我们采用了基于多任务的损失函数，它由鉴别器损失函数、身份保留损失函数和重建损失函数组成。我们的模型可以通过根据实际需求设置需要修改的面部属性数量来实现识别效用和人脸匿名之间的平衡，并提供不同设置下的结果。我们在两个公共基准 Celeb-A 和 VGG-Face2 数据集上进行的大量实验，证明了我们模型在不同人脸识别场景下的有效性。

**关键字**

隐私保护 人脸匿名 面部混淆

**1.介绍**

监控设备的大量使用已经记录了大量的面部图像数据，在此基础上许多基于人工智能的计算机视觉技术被用于大规模挖掘个人信息。因此，随着计算机视觉技术的巨大进步，有关隐私保护的问题越来越多。为了避免滥用隐私数据，一些限制性的法律法规，例如通用数据保护条例(GDPR)[1]，要求在使用隐私数据前需要征得个人对其隐私数据使用的许可。然而人脸图像数据的泄露在世界范围内仍然时有发生。此外，存储在数据库中的用户面部图像，即使没有暴露，但仍然容易受到第三方用户或应用的攻击而被窃取。因此，人脸图像的匿名化已成为许多人脸应用的关键步骤之一。

人脸图像匿名是一项具有挑战性的任务，它需要一个鲁棒的模型来修改原始人脸图像而不破坏现有的数据分布。现有方法[2-8]旨在去除所有识别信息，然后生成高度逼真的人脸。这些技术虽然降低了不必要的身份识别信息的隐私风险，但也破坏了人脸识别的便利性和安全性。因此，在监控数据库、驾驶执照数据库等其他应用程序中保留识别功能是十分必要的。

为了满足身份保护要求，Ross等人[9]引入了软生物特征隐私的概念，他们开发了一种在能够保留识别身份信息的前提下隐藏性别属性人脸混合方法。在他们后来的研究中，还引入了基于对抗性扰动的方法[10,11]来提高生成图像的视觉质量，使得生成的图像伪影较少。同样，Chhabera等人[12]已经采用对抗性扰动技术来实现保持身份的k属性匿名化。随后，PrivicyNet [13]被提出用来保护多属性（性别、种族和年龄）隐私，该模型可以提高生成的图像在跨多个人脸和属性分类器的泛化能力。不同的是，SensitiveNet[14]是从学习的嵌入空间中消除敏感信息，从而防止人脸识别系统的隐私泄露。尽管这些模型已经显示出它们在保持身份识别方面的能力，但由于生成的人脸图像相对于原始人脸图像几乎没有改变，因此侵犯用户隐私的风险仍然存在。对于实际使用，Li等人[15]提出了一种面部伪装模型，以保持监视场景下的识别有效性，可是这个模型需要一个参考人脸图像，而这将减少类内变化。因此，此技术仍然面临的挑战是如何确保人脸图像仅用于身份验证，而不会透露有关人脸的其他信息。

为了应对这一挑战，我们从人类认知的角度提出了一种新颖的面部匿名化模型，该模型可以在保留识别有效性的同时混淆视觉外观。我们首先引入身份感知模块来发现与个人身份识别密切相关的面部属性。然后，将原始人脸和发现的人脸属性指标输入人脸生成器，得到隐私保护下的人脸图像。具体来说，我们首先计算输入人脸的身份感知激活热图，得到支持身份预测的区域。接下来，我们将输入的人脸图像解析为五个部分（头发、眉毛、眼睛、鼻子和嘴唇），然后将每个部分与相应的语义属性关联起来。然后将各部分的得分相加，进而通过分数得到与身份无关的语义属性，即得到属性指标。之后，原始人脸和属性指示符一起输入到条件人脸生成器中，生成人脸匿名化图像。为了优化所提出的模型，我们采用了基于多任务的损失函数。它由鉴别器损失函数、重建损失函数和身份保持损失函数组成。我们的模型在两个公共基准Celeb-A[16]和VGG-Face 2[17]上进行了大量实验，证明了其在匿名视觉外观方面的有效性，并验证了其在不同人脸识别场景中的实用性。

最后，我们将开拓我们的面部匿名化模型的应用场景。在实际场景中，我们可以通过设计轻量级的嵌入式模型将面部匿名器部署到各种应用程序，包括监控、智能家居摄像头和机器人等场景。如图1所示，在将修改后的图像上传到数据库之前，这不会影响人脸图像的实用性并保护个人隐私。此外，我们的模型还可以在记录的人脸和真实人脸之间产生认知差距，而人类观察者可能无法关联这些差距。

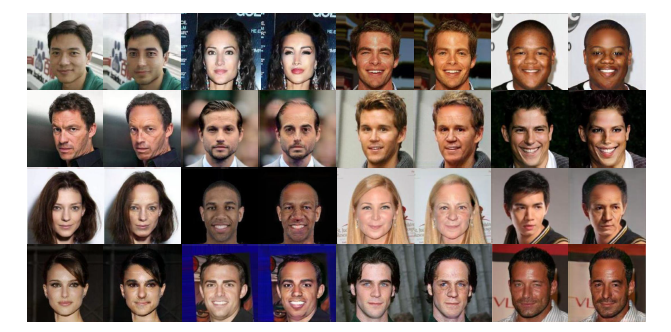


图1:我们提出的面部匿名化方法的说明。该方法能够自适应地发现与身份无关的视觉属性，并以这些视觉属性为条件生成保持隐私的人脸。新的人脸图像可以用于监控系统，因为它们的身份相关特征表示几乎没有改变。而观察者无法确定这两张脸是否代表同一个人，因为面部属性已经发生了变化。

**3.方法**

给定原始人脸图像 ，我们的方法旨在以自适应方式生成相应的保持身份的人脸图像。生成的人脸与原始人脸相比应该具有不同的视觉外观，但保留了与身份相关的特征。从认知的角度来看，人类通常对语义变化很敏感。因此，我们采用高感知敏感度的面部局部来评估视觉外观变化的程度。根据相关研究结果[41]，选取的面部部位是头发、眉毛、眼睛、鼻子和嘴唇。对于每个部分，我们选择相应的面部属性，即后退的发际线、浓密的眉毛、窄眼、大鼻子、大嘴唇。我们模型的流程图如图3所示。具体来说，我们首先通过CAM [42]计算身份感知激活热图，可用于定位与身份相关的面部部分。之后，我们使用人脸解析器[43]将人脸图像分成五部分并选择相应的人脸属性。然后，我们计算对面部属性进行排序的响应分数，以确定与身份无关的属性指标，其中c是属性的数量。接下来，将原始人脸和属性指示符 输入人脸生成器以获得最终的隐私保护人脸图像。

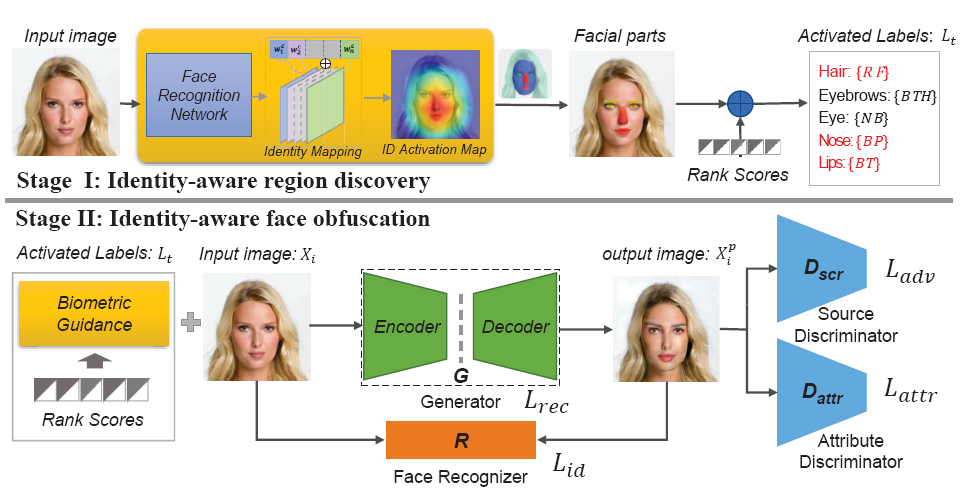


图3:提出的人脸匿名化框架。首先利用原始人脸图像计算身份识别类激活热图，然后通过对各部分激活得分排序得到身份无关属性指标。然后，将原始人脸和属性指示器送入条件人脸生成器，生成保护隐私的人脸。开发了一个多任务损失函数来实现端到端的训练过程，包括识别器丢失、属性分类器丢失和人脸识别丢失。

**3.1 确定身份相关的区域**

在本小节中，我们的目标是确定要保留哪些面部属性以及要混淆哪些面部属性。需要保留的属性代表与个人身份密切相关的关键面部特征，要混淆的属性代表所的人脸特征与个体的身份信息的关联十分松散。实际上，每个人的面部属性对于人脸识别器来说都是不同的，应该通过某种方式学习。因此，此模块的优点是它可以自适应地获取保留和混淆的面部属性。

给定所提出的框架，我们首先计算类激活图，它用于确定人脸识别中和判别相关的人脸部分。然而，由于感受野大，激活图不能直接用于发现身份感知区域。 因此，我们采用人脸解析器将人脸分割为人脸部分和人脸基础。面部部分是包含眼睛、眉毛、鼻子、嘴巴和头发的区域，它是从检测到的面部标志中获得的。之后，对于每个面部部分，我们计算激活图的像素平均值作为身份相关分数，然后我们按照分数的顺序对相应的面部部位进行排名 最后，每个面部部分都标有预定义的视觉属性，用作面部混淆模块中的属性指示符。

人脸的类激活图由CAM [42]模型生成。CAM 通常用于获取具有特定类型架构的图像分类CNN的响应热图其中全局平均池化卷积特征图直接馈入网络的softmax层。在这项工作中，我们首先用全局平均池化层替换主干网络的第二层到最后一层的全连接层，并使用softmax层计算特定身份标签的响应。具体来说，人脸识别模型的最终特征图表示为。其中W和H是特征图的宽度和高度，D是特征通道的数量。对于一个身份类c，预测分数是从特征图的全局平均池化结果中获得的：

(1)

其中是连接特征图的权重。身份类c的权重值可以直接从训练好的网络中获得。然后，身份c的热图被计算为特征通道的线性组合：

(2)

通过对进行上采样并将其添加到原始图像中，我们可以获得突出重要面部部位的定位图。 实验结果如图3所示。 第二列和第五列显示身份激活图，接下来的列是添加人脸分割的结果。这些肖像表明身份敏感的面部部位在不同的人之间并不一致：一些人的上部面部区域高度激活，而其他人则不然。

**3.2 人脸身份混淆**

在这项工作中，我们探索了一种新颖的面部混淆模型，该模型可以保留潜在的身份信息并显着改变视觉外观。要做到这些仅仅考虑图像翻译通常是不够的，我们需要找到一个可以同时用于外观变化和人脸识别的潜在代码。在这个模型中，应该实现面部特征匿名化和识别效用保留之间的平衡。引导人脸编辑的一种简单方法是通过基于StarGAN[33]的直接潜在代码优化，该方法已被广泛用于编辑人脸多属性图像。

该模块的结构如图3所示。它由四个子网络组成。生成器G用于生成具有目标标签Lt和原始人脸的合成人脸，训练源鉴别器Dsrc用于区分真实图像和合成图像，训练属性鉴别器Dattr用于预测生成的人脸属性。为了进一步提高生成的人脸的识别效果，我们还采用了辅助人脸识别器R。

具体地说,给定原始人脸和视觉属性指标,我们解决以下优化问题：

(3)

其中G为人脸生成器，D为保持合成图像的真实感，从而对人脸属性进行了精确修改。术语的概率用来表示是生成的图像是真实的还是合成的。生成器G和源鉴别器训练迭代优化,返回人脸图像输出的概率向量属于目标标签。生成器G试图生成能够被适当分类到目标标签向量的人脸图像。此外，利用损失项优化识别有效性的保留能力。

如上一节所述，人脸属性由保留属性和混淆属性组成。我们的模型经过训练以编辑混淆属性并维护由视觉属性指标获得的保留属性。对抗性损失指导模型最小化真实图像上的分布:

(4)

属性分类损失用于优化模型以获得所需的属性：

(5)

其中术语表示合成人脸属于类的概率。通过最小化这个目标，生成器G学习生成可以分类为所需属性的图像。属性鉴别器由标记的训练数据进行预训练，这是一个有监督的训练过程。

为了保留原始人脸图像中未选择属性的内容，我们将L1范数最小化作为重建损失:

(6)

其中表示原始属性标签。也就是说，在训练阶段使用G将生成的人脸图像重建为原始人脸图像。

最后，通过身份损失控制输入人脸与生成人脸的识别一致性:

(7)

其中是学习面部细节的感知损失，n是来自感知网络的卷积层集合，是来自第k层的激活。在我们的工作中，感知网络在人脸数据集上进行预处理，以确保生成的人脸包含识别特征。

我们模型的总目标函数定义为:

(8)

是一个平衡参数。上述方程中的前三项强制输出图像远离原始图像，而最后一项鼓励人脸识别器R判断两张人脸图像是同一个人。在我们的方法中,可以改变不同的迭代。首先，我们将其设置较小的值来让生成器真正改变人脸图像。然后,逐渐增大，从而把真实图像和生成图像的在R中得到的结果是一样的。

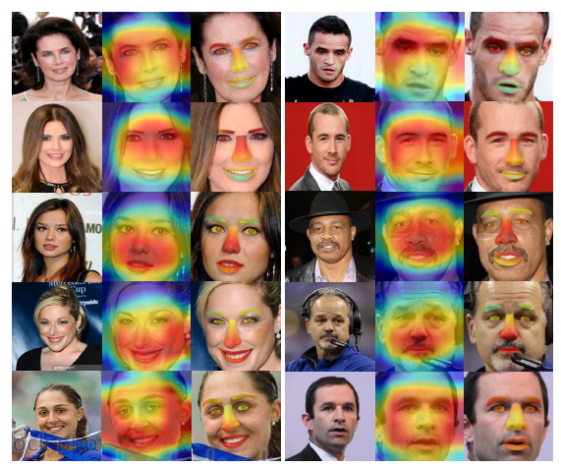


图 4：不同人的身份敏感面部部分。左列显示原始人脸，中列显示身份激活热图，右列显示 Identity Sensitive 的分割结果。对于人脸识别，身份敏感性面部部件的数量和位置因人而异。

在图5中，我们展示了人脸模糊模型生成的人脸序列。从每一行的第二个面开始，修改的面属性的数量减少。我们可以观察到面部外观有明显的变化。

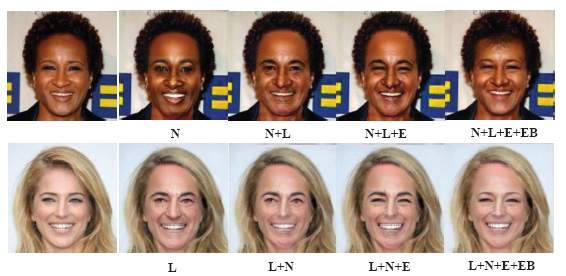


图 5：人脸混淆结果的说明。具有不同身份敏感部分的人的面部混淆结果。左列表示原始人脸，其余人脸由我们的模型生成，具有不同的属性指标。注释代表那些保留的 ID 敏感属性。（EB：眉毛，E：眼睛，N：鼻子，L：嘴唇）

**4 实验**

我们采用两个公共基准celebrity-a[16]和VGG-Face 2[17]数据集来评估我们的模型的有效性。实验有三种类型：对于面部外观匿名化评价，我们分析了生成图像的多样性和真实性的表现；为了保持身份信息，我们分析了生成的图像在三种人脸识别器上的性能；此外，还增加了一项用户研究，以验证该方法对人类观察者的有效性。

**4.1数据集**

Celeb-A[16]是最广泛使用的人脸属性识别数据集之一。它包含 10,177 个身份的 202,599 张图像。 每张图像有5个地标位置和40个属性注释。将原始人脸图像先裁剪为178\*178，然后resize为128\*128，其中我们随机选取2000张图片作为测试集，剩余的图片作为训练集。

VGG-Face2 [17]数据集包含 9,131个主题的331万张图像，每个主体平均 362.6张图像。图片是从谷歌图片搜索下载的，在姿势、年龄、光照、种族和职业（例如演员、运动员、政治家）方面有很大差异。具体来说，该数据集分为训练集和测试集，训练集由8,631个身份（3,141,890张图像）组成，测试集包含500个身份。我们从测试集中为每个身份随机选择一张图像。

**4.2 实现细节**

对于数据处理，我们遵循最近的模型[44,45]来裁剪和对齐人脸以获得归一化的人脸图像(112\*112)。对于人脸混淆模型，我们采用StarGAN[33]作为基本生成网络。在网络架构方面，我们采用了一个生成网络，其中包含两个用于下采样的卷积层、六个残差块和两个用于上采样的转置卷积层。该模型使用Adam进行训练，=0.5 和 = 0.999，我们在五次鉴别器更新后执行一次生成器更新。所有实验的批量大小设置为16。在CelebA的实验中，我们在前10个时期以 0.0001 的学习率训练模型，而学习率在接下来的10个时期线性衰减至0。对于身份保存任务，我们对三个人脸识别器执行黑盒攻击。具体来说，我们使用最近最先进的模型（Cosface[44]、Arcface[46]和VGG-Face2[17]）来评估人脸识别的性能。此外，我们探索了批量归一化和dropout层来训练强大的人脸识别器。

训练策略。在训练具有多个属性的人脸生成器时，我们使用人脸属性指标和原始人脸作为输入。同时，生成器应该学会忽略设置为零向量的未指定的视觉属性，并专注于明确的特定属性。除了输入属性指标的维度外，生成器的结构与 StarGAN 完全相同。另一方面，我们扩展了鉴别器的辅助分类器来预测所有数据集属性的概率分布。然后，我们在多任务学习设置中训练模型，其中鉴别器尝试学习两个数据集的鉴别特征，并最小化与属性相关的错误。在这些设置下，生成器学习控制两个数据集的所有视觉属性。

**4.3 评估标准**

**LPIPS距离**[47]用于衡量生成图像与原始图像之间的相似度。我们使用来自同一输入的随机采样输出对之间的平均LPIPS距离。LPIPS是根据图像深层特征之间的加权距离计算的。它已被证明与人类感知相似性有很好的相关性[47]。在[48]之后，我们对每个输入使用所有测试图像和相应的输出对，然后基于预训练的人脸识别模型提取它们的特征表示。

**FID距离**[49]用于计算真实图像的分布与生成的图像之间的距离。较低的 FID分数表示生成的图像质量更高，并且与真实图像相比更加多样化。因此，我们按照[49]中的实验设置来评估生成图像的真实性的性能。

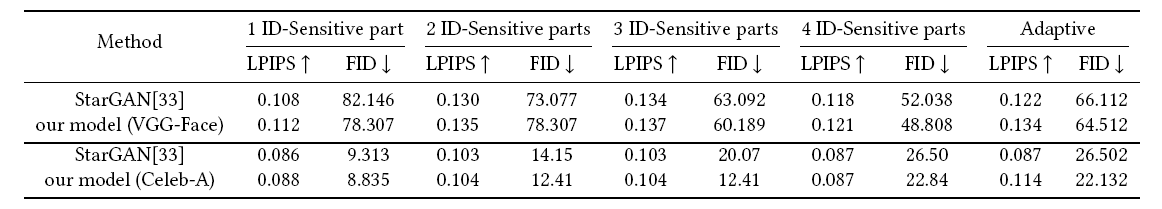
**人类偏好**。为了比较我们的模型生成的图像的真实性和忠实度，我们引入了人类感知研究。与[50]类似，我们将输入图像和我们的模型生成的图像呈现给工作人员。他们有无限的时间来选择哪些对是显着不同的。对于每次比较，我们随机选择不同的工人来回答，并使用平均分数作为真实情况。

**4.4 对外观匿名的评估**

我们模型的视觉外观匿名化在Celeb-A和VGG Face2数据集上进行了评估。在此任务中考虑修改不同数量的面部属性。我们在表2和图6中展示了定量和定性实验结果。对于定量评估，我们采用LPIPS和FID来评估生成图像的视觉变化和真实感。前者用于评估原始人脸和生成人脸之间的感知相似度。后者对视觉伪影很敏感，这可以表明生成的图像的真实性。我们的目标是将面部表现匿名化，在视觉质量和多样性方面，我们只需要一个实用的水平。因此，我们仅使用 StarGAN [33]作为比较基线。从表2中的实验结果，我们观察到我们的模型在两个基准的LPIPS和FID指标下都实现了更好的性能。同时，我们可以发现 LPIPS 和FID值随着修改的属性数量而变得更好。此外，在使用自适应机制生成保护隐私的人脸图像时，我们的模型仍然达到了预期的效果，并且没有降低生成图像的真实性和多样性。

图6显示了不同数量的面部属性修改的视觉实验结果。这表明我们的模型具有生成逼真图像的能力，与原始人脸相比，这些图像具有显着的视觉变化。一些典型的实验结果如图5所示。对于每一对，第一列是原始列，其他列由我们的框架生成。结果表明，我们提出的模型可以生成具有不同视觉外观但被识别为相同身份的新面孔。

表 2：我们的模型与StarGAN在CelebA和VGG Face 2数据集上对合成图像的真实性和多样性的定量比较。（1个ID 敏感部分表示一个面部属性已被混淆。）



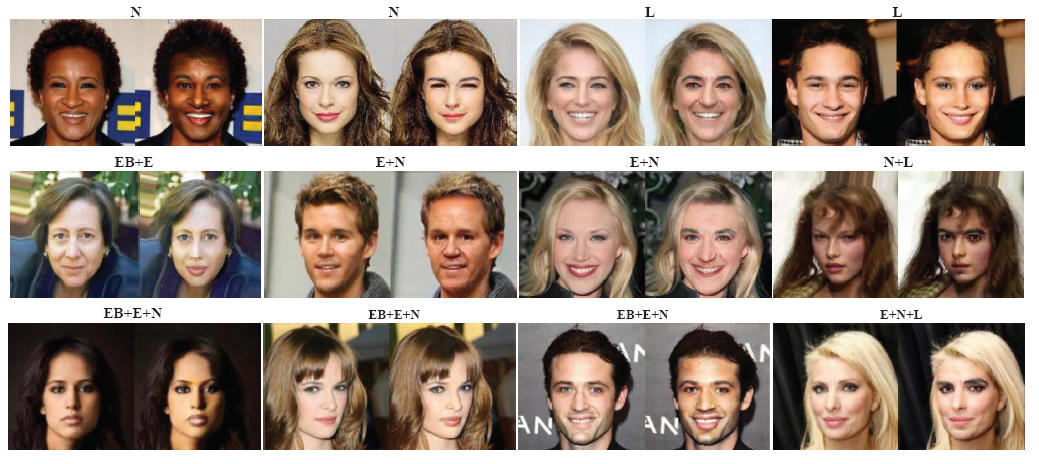


图6:不同id敏感部分编号的个人隐私保护结果，所有图像都是从测试集中采样的。左边的列显示未保护的面，右边的列显示由我们的方法生成的保护面部图像，尽管在外表上有很大的变化，但每一对仍然被认定为同一个人。注释表示保留的id敏感属性。(EB:眉毛，E:眼睛，N:鼻子，L:嘴唇)

**4.5 对身份保持的评估**

为了评估在匿名化视觉外观的同时保留身份的性能，我们对 Celeb-A 和 VGG Face2 数据集进行了一些实验。即在三个人脸识别器上进行了黑盒攻击的实验，三种广泛使用的人脸识别器：ArcFace[46]、CosFace[44]和VGG-Face 2 [17]，用于将原始人脸与匿名人脸进行匹配。具体来说，在 Celeb-A数据集上，我们随机选择了1,000个相同身份的人脸对和2, 000个不同身份的人脸对，以计算人脸识别器对人脸验证的平均准确率。而在VGG-Face 2数据集上，我们随机选择了500个相同身份的人脸对和1,000个不同身份的人脸对来开发测试列表。实验结果如表3和表4所示。我们的模型可以根据视觉属性的数量为每个输入人脸生成4张图像，同时该模型可以根据预定义的阈值自动寻找最佳人脸图像。因此，我们计算这5种情况下两个错误接受率(FAR)值的真实接受率(TAR)值。实验结果表明，我们的模型在三个人脸识别器上都达到了令人满意的平均人脸验证率，显示了不同人脸识别场景下的泛化能力。

表3:在FAR = 0.1和0.01时，不同视觉属性数量和不同人脸识别器在VGG face 2上保持身份人脸匿名化的表现，TAR(%)(1个id敏感部分表示有一个人脸属性被模糊化)。

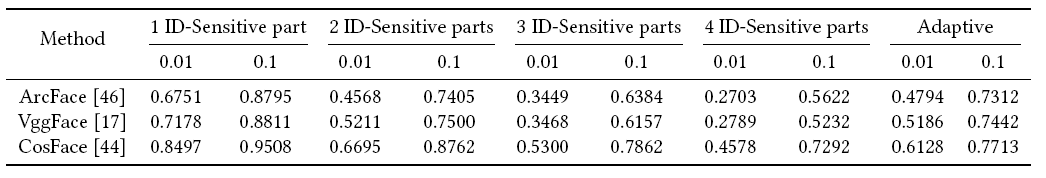


表4:在FAR = 0.1和0.01的情况下,以不同数量的视觉属性和明显的人脸识别器的身份保护的性能。(1个敏感的部分表明一个面部属性被混淆了。)

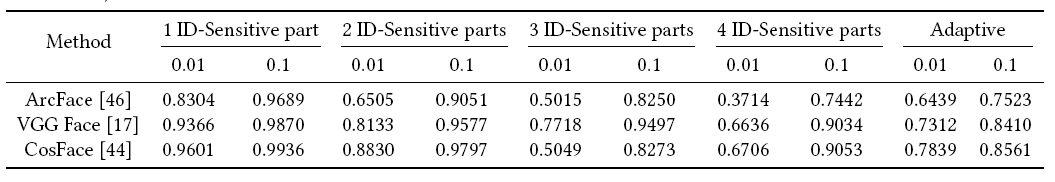
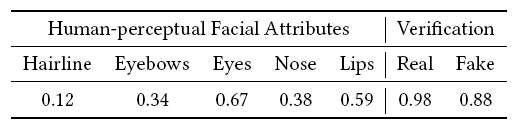


表5:用户研究人类感知面部属性及人脸验证实验结果



**4.6 用户研究**

为了评估我们的结果对于迷惑人类观察者的有效性，我们通过两个实验进行了用户研究。第一个实验旨在确定与身份相关的面部部分。为此，我们随机选择了200张不同身份的人脸图像，并为每张脸设置了五个选项（发际线、眉毛、眼睛、鼻子和嘴唇）。请用户从我们的选项中选择面部部位，他们可以进行单选或多项选择（不超过4个选择）。我们共收集了25名志愿者的5000个结果，并记录了不同面部部位的数量。表5中报告了每个面部属性的概率。结果表明，两个面部部分（眼睛和嘴唇）被高频选择，另外三个被低频选择。从人的角度来看，改变高频面部部分会使面部看起来像不同的人，而低频属性与身份的相关性较弱，尤其是发际线。因此，我们在合成过程中没有考虑修改发际线。

第二个实验旨在验证生成的人脸是否会混淆用户。具体来说，我们随机选择由100名男性和100名女性组成的200个图像对，属于同一个人和不属于同一个人的图像对数量相等。然后，我们从25名志愿者那里收集了总共5000条结果。 最后，我们计算人脸验证的平均准确率来衡量视觉外观匿名化的能力，如表5所示。结果表明，我们的方法生成的匿名人脸可以显著降低人类的识别率。

**5 总结**

在本文中，我们开发了一种新的框架来保护监控系统中人脸图像的隐私。 我们引入了一种新颖的面部匿名化模型，它将StarGAN的强大生成能力与发现的非身份相关的面部属性指标相结合。我们已经证明，我们的模型能够在身份保留的情况下生成面部图像的广泛外观变化。实验结果表明，我们的方法可以保留不同人脸识别器的识别效用，并有效地匿名化面部外观。我们还证明了我们的方法提供了细粒度的编辑控件，例如指定所需的属性，例如大鼻子。未来，我们将探索我们的模型保护其他面部属性的能力，例如种族、年龄、情感等。

**致谢**

国家重点研发计划(2018aaa0100601)、国家自然科学基金项目(62025604,62072454,U1936210)、北京航空航天大学虚拟现实技术与系统国家重点实验室开放项目(vrlab2021c06)、北京市自然科学基金项目(4202084)资助

**引用**

[1] Razvan Viorescu et al. 2018 reform of eu data protection rules. European Journal of Law and Public Administration, 4(2):27–39, 2017.

[2] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In Proceedings of the IEEE International Conference on Computer Vision, pages 9378–9387, 2019.

[3] Hanxiang Hao, David Güera, Amy R Reibman, and Edward J Delp. A utilitypreserving gan for face obscuration. arXiv preprint arXiv:1906.11979, 2019.

[4] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[5] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In European Conference on Computer Vision, pages 727–743, 2020.

[6] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang.

Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[7] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5447–5456, 2020.

[8] Shuhui Yang, Han Xue, Jun Ling, Li Song, and Rong Xie. Deep face swapping via cross-identity adversarial training. In International Conference on Multimedia Modeling, pages 74–86, 2021.

[9] Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In European Conference on Computer Vision, pages 682–696. Springer, 2014.

[10] Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In 2017 IEEE International joint conference on biometrics (IJCB), pages 564–573. IEEE, 2017.

[11] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semiadversarial networks: Convolutional autoencoders for imparting privacy to face images. In 2018 International Conference on Biometrics (ICB), pages 82–89. IEEE, 2018.

[12] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 656–662, 2018.

[13] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: semi-adversarial networks for multi-attribute face privacy. IEEE Transactions on Image Processing,29:9400–9412, 2020.

[14] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana.Sensitivenets: Learning agnostic representations with application to face images.IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[15] Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. Learning disentangled representations for identity preserving surveillance face camouflage. In 25th International Conference on Pattern Recognition, 2020.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[17] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pages 67–74, 2018.

[18] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. IEEE transactions on Knowledge and Data Engineering,17(2):232–243, 2005.

[19] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1491–1500. IEEE, 2017.

[20] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In International Symposium on Visual Computing, pages 565–578. Springer, 2019.

[21] Carman Gerard Neustaedter and Saul Greenberg. Balancing privacy and awareness in home media spaces. Citeseer, 2003.

[22] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In Proceedings of the 2000 ACM conference on Computer supported cooperative work, pages 1–10, 2000.

[23] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Modelbased face de-identification. In 2006 Conference on computer vision and pattern

recognition workshop (CVPRW’06), pages 161–161. IEEE, 2006.

[24] Liang Du, Meng Yi, Erik Blasch, and Haibin Ling. Garp-face: Balancing privacy protection and utility preservation in face de-identification. In IEEE International Joint Conference on Biometrics, pages 1–8. IEEE, 2014.

[25] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face deidentification.In 2015 International conference on biometrics (ICB), pages 278–285.IEEE, 2015.

[26] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models.In 29th {USENIX} Security Symposium ({USENIX} Security 20), pages 1589–1604,2020.

[27] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5050–5059,2018.

[28] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. arXiv preprint arXiv:2005.10329, 2020.

[29] Yunqian Wen, Li Song, Bo Liu, Ming Ding, and Rong Xie. Identitydp: Differential private identification protection for face images. arXiv preprint arXiv:2103.01745,2021.

[30] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Unsupervised privacy-enhancement of face representations using similarity-sensitive

noise transformations. Applied Intelligence, 49(8):3043–3060, 2019.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-toimage translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 2223–2232, 2017.

[32] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In Advances in neural information processing systems, pages 700–708, 2017.

[33] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and

Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer

vision and pattern recognition, pages 8789–8797, 2018.

[34] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen.

Attgan: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11):5464–5478, 2019.

[35] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. arXiv preprint arXiv:2103.01456, 2021.

[36] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1415–1424, 2017.

[37] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. Sdit: Scalable and diverse cross-domain image translation. In Proceedings of the 27th ACM International Conference on Multimedia, pages 1267–1276, 2019.

[38] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan:Learning a symmetry three-player gan for identity-preserving face synthesis.In Proceedings of the IEEE conference on computer vision and pattern recognition,pages 821–830, 2018.

[39] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards

diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5549–5558, 2020.

[40] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[41] Galit Yovel Naphtali Abudarham, Lior Shkiller. Critical features for face recognition.Cognition, 182, 2019.

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba.Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016.

[43] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In 2012 IEEE Conference on Computer Vision and Pattern Recognition,pages 2480–2487. IEEE, 2012.

[44] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE conference on computer vision and pattern

recognition, pages 5265–5274, 2018.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[46] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In CVPR, 2019.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and OliverWang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595,2018.

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, OliverWang,and Eli Shechtman. Toward multimodal image-to-image translation. arXiv preprint arXiv:1711.11586, 2017.

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv preprint arXiv:1706.08500, 2017.

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro.High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8798–8807, 2018.