人脸跨域合成技术研究进展

刘琦[1],吴昊展[2],谢添鑫[2],韩琥[3]

1. 河南警察学院网络安全系,郑州 450000
2. 中国科学院大学,北京 100049
3. 中国科学院计算技术研究所,北京 100190

摘要　总结了人脸跨域合成技术的起源、任务类型与难点、技术发展与挑战、潜在应用与问题等,从自监督与弱监督跨域合成、基于预训练大模型跨域合成、基于跨域合成隐私保护3个方面探讨了人脸跨域合成技术未来发展趋势与挑战。

关键词　人脸跨域合成;虚拟人;生成对抗网络;神经渲染;扩散模型

人脸跨域合成是指从一个域的人脸图像合成另一个域的人脸图像,例如,从人脸照片合成人脸画像,从悲伤的人脸合成喜悦的人脸,从侧面人脸合成正面人脸等。人脸跨域合成既可以是对整个人脸的合成,也可以是对局部人脸的合成。人脸跨域合成在影视制作、社交通信、互联网媒体等诸多领域都有广泛的应用。

早期的人脸跨域合成主要面向跨姿态、跨光照的人脸合成,通常基于图像处理技术或三维人脸重构技术进行。基于图像处理技术的方法可以通过图像滤波、图像补全等实现人脸图像超辨率、人脸图像去遮挡的功能,但能实现的任务有限,且难以进行大幅度的人脸变换。基于三维人脸重构的方法是先从二维人脸图像恢复三维人脸的形状和纹理等信息,进而基于三维人脸模型进行跨域人脸图像合成。Blanz 和 Vettert[1]在1999年提出的三维人脸形变模型(3D morphable model,3DMM)是三维人脸重建代表性方法之一。该方法基于三维人脸数据进行主成分分析所获得的三维人脸形状和纹理主成分,建立线性表示模型以拟合新的人脸三维形状和纹理,进而通过将三维人脸的形状和纹理重渲染并逼近输入二维人脸图像,实现对线性模型参数的迭代优化[2]。三维人脸形变模型延伸出一系列工作,并随着深度学习技术飞速发展,形成了利用深度模型预测线性模型参数的新路线[3],取得了比传统迭代优化更好的效果。跨姿态、跨光照人脸合成技术可以应用于人脸识别等领域,降低因人脸图像光照和姿态差异导致的识别错误率。Tang 和 Wang[4]在2003年针对画像—照片跨模态合成,提出了基于形状和纹理特征变换的跨模态合成方法。该方法假设画像和照片之间的差异可以通过一个线性变换来近似,从而基于主成分分析建立各模态

收稿日期:2023-04-23;修回日期:2023-07-13
基金项目:河南省科技厅科技攻关项目(222102210089)
作者简介:刘琦,副教授,研究方向为信息技术及应用,电子信箱:569797767@qq.com
引用格式:刘琦, 吴昊展, 谢添鑫, 等. 人脸跨域合成技术研究进展[J]. 科技导报, 2023, 41(16): 113-123; doi: 10.3981/j.issn.1000-7857.2023.16.010

图像的线性表示模型,实现跨模态人脸图像合成。跨模态人脸合成可以应用于异质人脸识别,如画像—照片异质人脸识别、可见光-近红外异质人脸识别等领域,降低因图像模态差异导致的人脸识别错误率。以面部动作合成为代表的跨域人脸合成因其在影视制作等领域的潜在应用价值,也成为早期人脸跨域合成的重要研究方向之一[5-6]。早期的方法面向普遍使用基于标记点的面部表情捕捉系统从演员面部提取行为表情信息,进而驱动预先准备好的三维人脸模型产生相应的行为表情。这些技术在《指环王》(《The Lord of the Rings》)、《阿凡达》(《Avatar》)等大量影视作品中获得广泛应用。近年来,随着深度学习技术的飞速发展,特别是生成对抗网络(generative adversarial model, GAN)[7]、风格迁移模型(StyleGAN)[8]、神经辐射场模型(neural radiance field,NeRF)[9]和扩散模型(diffusion model)[10]等技术的进步,使得人脸跨域合成可以摆脱三维人脸重建技术、线性表示模型和基于标记点的面部表情捕捉系统的限制,以端到端的方式进行。这些新技术有效降低了人脸跨域合成的成本和复杂度,提升了人脸跨域合成的质量。以面向人脸属性编辑的跨域合成为例,目前最好的端到端方法,在开放场景数据集CelebA-HQ上,其合成人脸图像的FID(frechet inception distance)误差已经降低到10以内,能较好地满足一些实际应用的需求。

在过去几十年中,人脸跨域合成取得了巨大的进步,其技术路线也从原来的每类跨域合成任务单独建立一个模型发展为相对统一的模型。然而,现有的人脸跨域合成技术仍然面临诸多挑战,限制了其在更开放、更广泛场景的应用:(1) 传统的基于3D人脸重建的跨域合成方法在大姿态、有遮挡情况下,因为重建误差,会导致跨域合成效果变差;(2) 基于标记点的动作单元捕捉系统虽然能有效提升人脸跨域合成对姿态变化的鲁棒性,但标记点的使用限制了其应用便利性;(3) 基于端到端深度学习的人脸跨域合成虽然降低了人脸跨域合成的复杂度,但其跨域合成的分辨率仍然受到限制,难以直接生成2K或4K高分辨率图像。为此,本文对人脸跨域合成的常见任务、发展历程、技术原理、典型应用等进行总结,进而对其未来发展趋势和潜在的挑战进行探讨。

1 人脸图像域及跨域合成任务

为了更加系统地呈现人脸跨域合成技术,总结了常见的人脸图像域和跨域合成任务,并对不同任务的特点进行简要介绍。如图1所示,人脸图像域的划分与任务密切相关,常见的跨域合成任务中大致涉及的图像域类型,可以分为不同图像模态、不同光照姿态、不同行为属性、不同图像质量、不同干扰噪声等。基于这些图像域的定义,有一些对应的跨域合成任务,例如:跨模态合成、跨光照姿态合成、跨属性合成、人脸重现生成、跨质量合成、对抗样本生成等。

1.1 常见人脸图像域

1) 不同图像模态域。人脸图像的模态是指因不同图像获取方式而得到的不同图像,例如,可见光图像、近红外图像、深度图像、热红外图像等。不同模态的图像蕴含的人脸信息往往不同,如可见光人脸图像富含纹理和色彩信息,但容易受光照的影

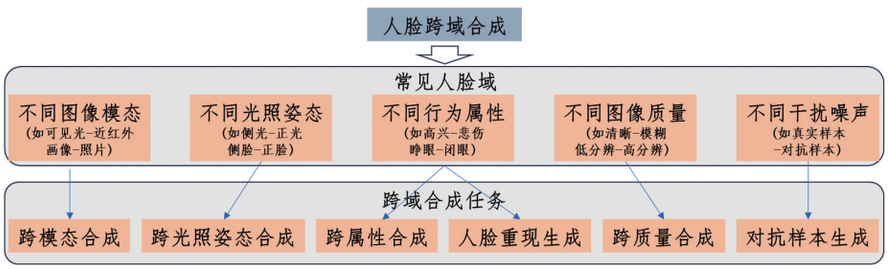


图1 常见的人脸图像域及跨域合成任务

响;近红外和深度图像受光照影响小,但缺乏丰富的纹理和色彩信息。此外,人脸图像的模态也包含人脸画像(或草图、素描、油画)等形式的人脸图像。这些不同的图像模态可以视为不同的人脸图像域。

2)不同光照姿态域。人脸图像成像过程中,往往因人脸角度、相机角度以及光照变化,形成不同光照、姿态的人脸图像,如正面光照和侧面光照,正面人脸和侧面人脸。不同光照或姿态的人脸图像之间往往存在较大的表观差异,如正面人脸中通常双目可见,而侧面人脸中可能仅有单目可见,因而其反映的人脸信息也不同,不同光照或姿态可以视为不同的人脸图像域。

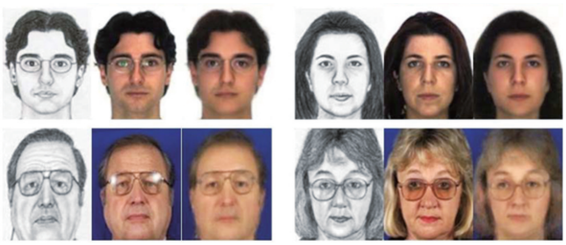
3)不同行为属性域。人脸是非刚性的物体,会随着表情、动作的变化而形成不同的人脸图像。例如:在表情方面,人脸包含7种常见的表情变化,快乐、悲伤、愤怒、恐惧、惊讶、蔑视和厌恶;在行为方面,人脸会呈现张嘴、闭眼等动作。此外,人脸的五官、年龄、性别、妆容等特征也多种多样。每种表情、动作、属性等都涵盖了一类人脸图像,因此可以视为不同的行为属性域。

4)不同图像质量域。开放场景下,因摄像机分辨率差异、人脸距离远近、人脸运动状态等因素,人脸图像成像质量参差不齐。例如,在理想光照条件下,人脸图像清晰明亮;反之,人脸图像模糊暗淡。在近距离条件下,人脸图像分辨率高;反之,人脸图像分辨率低。此外,人脸图像中还常存在遮挡与自遮挡问题,导致部分区域不可见。不同分辨率、不同模糊程度、不同遮挡程度、不同噪声程度的人脸图像类型分别构成了不同的图像质量域。

5)不同干扰噪声域。干扰噪声域和图像成像中的噪声有所不同,干扰噪声主要是指人为加入的噪声,以改变图像原有的类别信息,加入干扰噪声后,图像的视觉效果不一定会有明显下降,因此这里将干扰噪声域单独作为一类。目前的应用中主要包含有对抗干扰噪声的图像和无对抗干扰噪声的图像2类。这2个域的图像可以相互转换,例如,在隐私包含中,可能需要将无干扰噪声的图像添加干扰噪声,使得现有的人脸分析模型无法从中提取用户的隐私信息;另一方面,也可以设计干扰噪声的图像检测和净化模型,使得输入人脸分析模型的图像是经过去干扰噪声处理的,从而保护人脸分析系统免受欺骗,避免他人的人脸图像被滥用冒用。

1.2　人脸跨域合成任务

1)人脸跨模态合成。不同模态域的人脸图像之间存在人脸表观信息的差异,往往需要进行跨模态人脸图像合成,实现对不同模态的转换和统一。人脸跨模态合成通常是一组对偶问题,即可以进行双向跨域合成,如从人脸画像合成人脸照片和从人脸照片合成人脸画像。图2展示了文献[11]从人脸画像到照片的跨模态合成例子,其目标是让合成的人脸照片尽可能接近真实人脸照片。

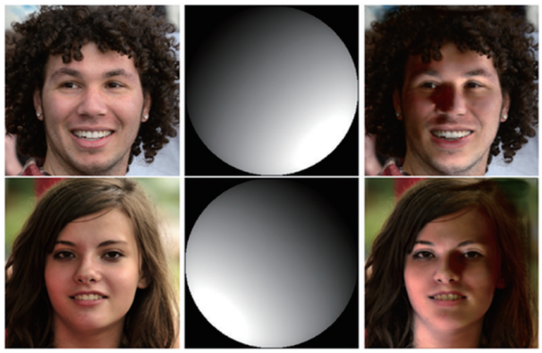


(a)(d)为人脸的画像;(b)(e)为真实的人脸照片;(c)(f)为跨域合成的人脸照片

图2　人脸画像到照片的跨域合成

人脸跨模态合成任务的难点是实际场景中往往缺乏大规模成对数据进行有监督建模,因此难以采用像素级约束来提升人脸跨模态合成的质量。此外,如何有效保持原人脸图像的人物身份不变也是人脸跨模态合成的一个挑战性问题[11]。

2)人脸跨光照姿态合成。不同光照姿态域的人脸图像之间存在较大表观差异,这些差异往往导致进行后续人脸分析和识别任务时性能下降;因此,很多应用场景中,需要进行跨光照姿态的人脸合成,降低不同光照或姿态图像之间的差异,提升任务的性能。图3展示了文献[12]跨光照人脸合成的示例,其目标为原光照人脸图像经跨光照合成的人脸图像尽可能接近目标光照类型。人脸跨光照姿态合成中,除了利用图3(b)所示的球面谐波光照输入外,还可利用另一幅人脸图像中的光照作为目标光照输入,进行基于样例的跨光照人脸合成。

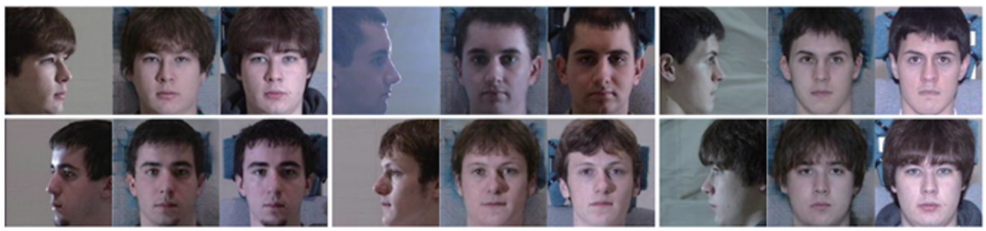


(a) 原光照人脸图像 (b) 目标光照类型 (c) 经跨光照合成的人脸图像

图3 人脸跨光照合成示例

图4展示了文献[13]的跨姿态人脸合成示例,共包含3组,每组的第1列展示了原姿态人脸图像,即60°、75°、90°的侧脸图像;第2列展示了跨姿态合成的正面人脸图像,其目标是尽可能接近每组第3列所示真实的正面人脸图像。

人脸跨光照姿态合成同样面临缺乏大规模成对数据的挑战,难以仅通过强监督学习建立脸跨光照姿态合成模型,以及利用逐像素损失函数来约束跨光照姿态合成的效果。人脸跨光照姿态合成中,同样期望合成图像能有效保持原人脸图像的身份不变。此外,跨光照姿态合成在处理复杂光照和姿态时仍面临很大的挑战,容易出现合成真实度不高的问题。



(a) ±60° (b) ±75° (c) ±90°

图4 人脸跨姿态合成示例

3)人脸跨行为属性合成。不同的行为属性域表达了人脸不同的表情、行为、五官特征。在影视制作等应用中,经常需要改变人脸原本的行为表情,合成出具有新行为表情的人脸图像。在合成新的人脸行为与表情过程中,既可以通过建模并控制影响行为表情的变量来合成新的行为表情人脸图像,也可以将另一个人脸的行为表情作为参考目标进行行为表情迁移(又称为人脸重现)。图5展示了文献[14]人脸跨行为属性合成示例。

(a) (b) (c) (d)

(a)为原始的人脸图像;(b)~(d)分别为通过跨行为属性人脸合成的年龄老化、戴眼镜、变性别的图像

图5 人脸跨行为属性合成示例

因为人脸行为属性的多样性,而且很难在同一时刻获取同一人脸不同行为属性的图像,跨行为属性合成任务也缺乏大规模成对数据,因而难以进行大规模有监督的跨行为属性人脸合成建模。此外,人脸不同的行为属性往往耦合在一起,如何在对某种行为属性进行跨域合成时,保持其他行为属性不变也是跨行为属性合成的挑战性问题。

4)人脸重现合成。人脸重现是人脸跨行为属性合成的一种,即将另一个人脸的行为表情作为目标进行行为表情迁移。近年来人脸重现合成技术飞速发展,达到了以假乱真的效果,被恶意用于伪造政治家、企业家等公众人物的演讲,引发了国际社会的广泛关注。人脸重现合成是指将源域人脸的面部行为表情迁移到目标人脸上,使目标人脸具有与源域人脸一致的行为和表情。图6展示了文献[15]人脸重现合成示例,人脸重现合成就是要把图6(b)中的人脸行为表情,迁移到图6(a)所示的

被驱动的目标人脸中,使图6(a)中的人脸动起来,达到接近真人在说话的效果。

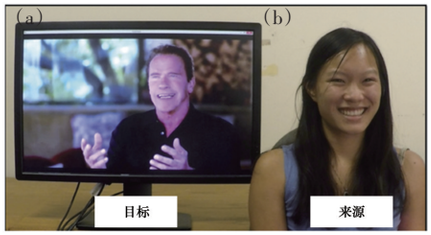


图6　人脸重现合成示例

人脸重现合成的难点在于如何将源域和目标域人脸中的行为表情与其他信息进行解耦,从而避免人脸重现合成的图像出现身份混淆的问题。此外,人脸重现合成往往缺乏前述几个任务中的金标准图像,用于定量地评估人脸重现合成的效果。

5)人脸跨质量合成。不同质量的人脸图像给诸多人脸分析任务带来挑战,因此在人脸识别等应用场景中,往往需要将低分辨率人脸图像转换为高分辨率人脸图像,将高噪声人脸图像转换为低噪声人脸图像,即跨质量人脸合成,从而提升下游任务的鲁棒性。图7展示了文献[16]从低分辨率到高分辨率的跨质量人脸合成示例,低分辨人脸图像图7(a)通过跨合成得到的高分辨率人脸图像图7(b),目标是接近图7(c)中的真实高分辨率图像。图8展示了文献[17]从遮挡人脸到无遮挡人脸的跨域合成示例。

人脸跨质量合成任务也普遍面临缺乏成对数据的难题,现有的研究中通常利用模拟数据进行模型训练和验证,如将高分辨率人脸图像降低为低分辨率人脸图像。然而,监控、刑侦等真实场景中,基于模拟数据训练的跨质量合成模型可能泛化能力不好,而且往往没有真实图像用于定量的评价合成的效果。

图7　人脸跨分辨率合成示例

(a)有遮挡的人脸图像

(b)去掉遮挡合成的人脸图像

图8　人脸去遮挡合成示例

6)人脸对抗样本生成。人脸对抗样本生成可以看作是一种广义的不同图像质量域的跨域合成。由于近年来人脸对抗样本生成技术飞速发展,且存在被恶意使用的风险,成为广受关注的热点之一。具体来说,对抗样本生成是指通过对抗学习的方式给图像加入轻微的扰动噪声,导致原本可以被正确识别的图像被误识别为其他类。图9展示了文献[18]对抗样本生成示例,图9(a)为一张熊猫图片,能被分类模型以57.7%的置信度分类正确;当加入图9(b)所示噪声后,得到对抗样本图像图9(c)被分类模型以99.3%的置信度错误地分类为长臂猿。类似的技术同样可以用于人脸对抗样本生成。

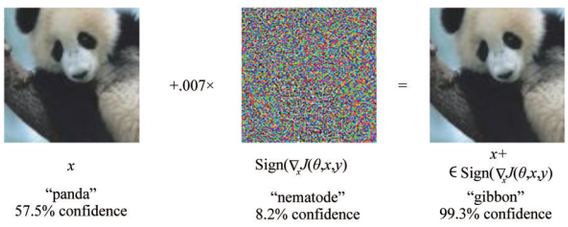


图9　人脸对抗样本生成示例

人脸对抗样本生成的挑战主要包括合成样本的自然度和真实度,以及面向未知识别模型(黑盒攻击)情况下的对抗样本的泛化能力。鉴于人脸对抗样本生成技术有被恶意使用的风险,因此如何通过人脸对抗样本生成,发现并提升模型的鲁棒性也是这一方向的重要研究问题。

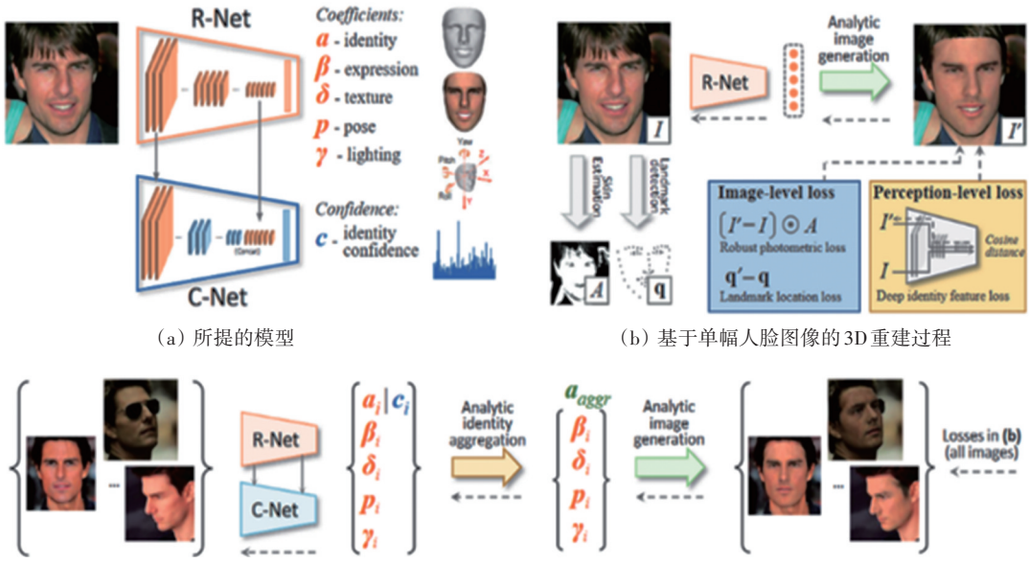
2 人脸跨域合成技术

不同的人脸跨域合成任务往往对应不同的人脸跨域合成技术,现有的人脸跨域合成技术可以大致归为5类:基于图像处理的方法、基于三维重建的方法、基于子空间的方法、基于生成模型的方法、基于神经渲染的方法。

1)基于图像处理的方法。早期的人脸跨域合成技术大多采用基于图像处理的方法,例如图像滤波、亮度调整、图像补全等图像处理技术,实现光照转换、图像锐化、图像超分辨、图像修复等功能。这些方法的特点是一般不需要通过训练数据进行模型训练,就能直接在任意一张人脸图像上进行操作,但对操作者的经验有比较高的要求。此外,受限于这些方法对全局人脸信息建模能力的不足,基于图像处理的方法很难进行大幅度和大范围的人脸跨域合成,否则合成的人脸图像中往往包含严重的图像瑕疵,图像的真实度也不高。

2)基于三维重建的方法。基于三维重建的跨域合成方法通常需要先从单张或多张人脸彩色图像进行三维人脸重建,得到人脸的三维形状、纹理、表情、姿态等信息,然后对相应的信息进行修改,实现对人脸图像的跨姿态、跨光照、跨表情合成。图10展示了文献[19]基于深度学习的人脸三维重建模型。其通过深度卷积神经网络,预测人脸的身份、表情、纹理、姿态、光照系数,进而可以利用更改其中相应的系数,并重新渲染得到二维人脸图像,实现跨姿态、跨光照、跨表情合成。该类跨域合成人脸方法的主要挑战在于用人脸三维重建本身是一个病态问题(ill-posed problem),人脸三维重建的准确度直接决定了人脸跨域合成的效果。此外,该类方法从三维人脸重渲染二维人脸过程中,也依赖于渲染器的能力。

3)基于子空间的方法。基于子空间的方法是前深度学习时代人脸跨域合成的主要方法,如人脸画像到照片合成、人脸低分辨率到高分辨率合成、人脸遮挡到无遮挡合成等。基于子空间的方法通常假设每个图像域中的人脸图像或人脸的局部区域可以被域中的其他人脸或人脸区域线性表示,从而在每个图像域建立一个线性子空间模型(如全局线性模型、局部线性模型),进而通过分别求解2个域中的线性模型系数,实现人脸图像从一个域到另一个域的合成。图11展示了文献[20]线性子空间模型人脸画像与照片跨域合成,主要基于主成分分



(a)所提的模型 (b)基于单幅人脸图像的3D重建过程

(c)基于多幅人脸图像的3D重建过程

图10 基于深度学习的三维人脸重建

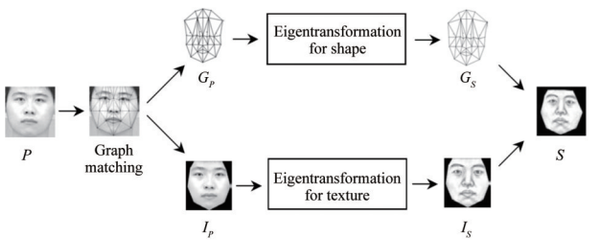
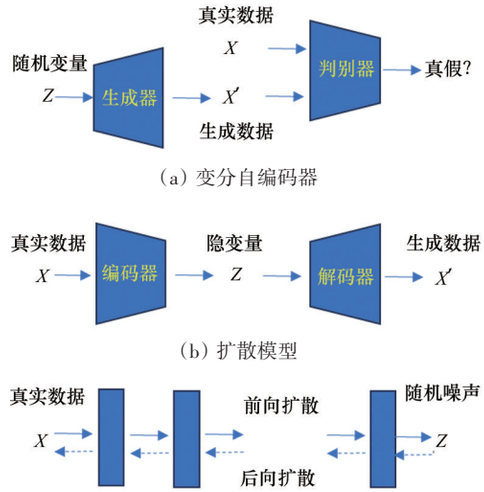


图11　基于线性子空间模型人脸画像与照片跨域合成

析建立形状特征变换和纹理特征变换。基于子空间的人脸跨域合成方法通常对数据规模的要求不高,具有较高的算法效率,但通常依赖于人脸线性表示的假设,在人脸具有复杂表观变化时,这些假设可能会失效,影响跨域合成的效果。

4)基于生成模型的方法。生成模型能够根据观测数据学习样本的分布,进而随机生成与观测数据同分布的数据,代表性的生成模型包括:生成对抗网络(generative adversarial net,GAN)[7]、变分自编码器(variational autoencoder,VAE)[21]、扩散模型(diffusion model)[22]。图12展示了生成对抗网络、变分自编码器和扩散模型的基本方法框架。生成对抗网络主要包括1个生成器和1个判别器,其中,生成器通过真实图像分布让自身生成的图像尽可能接近真实图像,以骗过判别器;而判别器则尽力区分生成图像和真实图像;两者通过不断对抗博弈,最终达到平衡。变分自编码器同生成对抗网络类似,也包含编码器和解码器,其中编码器将输入空间映射到与变分分布的参数相对应的潜空间;解码器则将潜空间映射回输入空间,从而能生成更多同分布的数据。扩散模型包含一个可以通过马尔



(a)变分自编码器

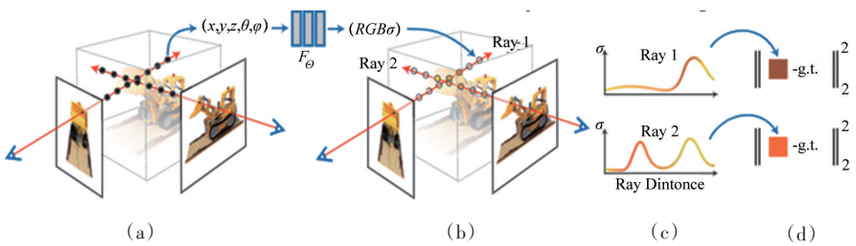
(b)扩散模型

(c)基本框架

图12　生成对抗网络

可夫链建模为图像逐渐添加高斯噪声的前向扩散过程和一个通过神经网络学习为图像逐渐去除噪声的后向扩散过程。

5)基于神经渲染的方法。传统的渲染方法主要是栅格化与光线追踪等,这些渲染方法需要输入渲染对象和场景的几何形状、材质属性(如反射率、透明度)、环境光照等信息,但是精确获取真实场景的所有物理参数难度很大,这限制了传统渲染方法的真实性。神经渲染则结合了经典计算机图形学与深度学习方法,仅需输入少量的场景信息,用神经网络学习从场景信息到照片的渲染过程,即可实现真实度较高的渲染结果,这是朝着合成具有真实感图像迈出的一大步。图13展示了文献[9]的神经辐射场(neural radiance field,NeRF)模型。该方法



(a)沿着相机光线以5D坐标作为MLP的输入;(b)MLP输出颜色和体积密度;(c)使用题绘制技术的渲染;(d)使用合成图像和真实图像的差异来优化场景表示

图13　神经辐射场模型示意

将物体建模成一个连续的5D辐射场,并隐式地存储于神经网络中。只需利用多角度的2D图像,就可以训练得到一个神经辐射场模型,根据这个模型可以渲染出任意视角下的清晰照片,实现跨视角合成。基于神经渲染的跨域合成方法的优点是生成的三维模型质量高、逼真度高,可以在任意视角和距离进行跨域生成,且不需要对输入图像进行特定处理或标注。

3 人脸跨域合成技术的应用

人脸跨域合成技术已经应用于诸多领域,如虚拟人物、美颜美化、司法刑侦、模型增强等。这些应用提高了研究人员对人脸跨域合成技术的关注度,有助于推动人脸跨域合成技术的进一步发展。

1)虚拟人物。基于人脸跨域合成的虚拟人物生成在影视制作、社交通信等领域应用广泛。在影视制作领域,早在2009年上映的科幻电影《阿凡达》中,所有的演员都穿着带有标记点的动作捕捉服,头戴配备摄像头的特殊装置,进行从人脸到人体的全方位动作捕捉,将真人动作、表情迁移到虚拟人物上,实现虚拟人物合成。在电影《速度与激情7》(《Fast & Furious 7》)拍摄期间,此前担任主演的保罗·沃克(Paul Walker)意外去世,影视制作团队通过用替身扮演这个角色,之后利用人脸跨域合成技术将保罗·沃克的面部图像替换上去,实现了保罗·沃克的重现。在社交通信领域,很多视频通信工具都支持基于人脸视频驱动虚拟的2D或3D Avatar头像的功能。在虚拟教学中,可运用人脸跨域合成技术,将授课人的人脸换为名师的人脸,更好地吸引学生的学习兴趣。此外,虚拟人物在新闻播报、智能客服等领域也都有越来越多的应用。

2)美颜美化。在人像摄影、视频直播等应用中,存在大量美颜美化需求,例如:皮肤美白度不高,面部存在皱纹、斑痣,面部偏胖等。基于人脸跨域合成技术,可以实现诸如人像皮肤美白、皮肤磨皮、祛痘祛斑、瘦脸瘦鼻、智能上妆、美发美牙等功能。目前,常用的手机摄像软件、图像处理软件(如Photoshop、美图秀秀),视频聊天工具(如QQ)、视频分享软件(如抖音、快手)等,都大量采用人脸跨域合成技术进行人像美颜美化来提升人脸图像的视觉效果。

3)司法刑侦。早期在司法刑侦领域,经常存在难以获取犯罪嫌疑人照片的情形,此时往往需要根据目击证人的描述建立嫌疑人的人脸画像。人脸画像一方面可以用于发布在媒体上通缉嫌疑人,另一方面可以用于与已知人脸库的人脸比对。因为人脸画像与照片之间的差异,可以通过人脸画像—照片合成技术,将画像转换为照片,进而与人脸库中的照片进行比对,提高人脸识别的精度[23]。例如,根据公开报道,西安电子科技大学等为警方联合研发的"智慧之眼"系统,实现了将嫌疑人模拟画像合成为高清照片的功能,也可以将警方人脸照片合成为人脸画像;然后,分别在人脸照片库和人脸画像库中进行人脸比对,以快速缩小侦查范围,锁定嫌疑人。

4)模型增强。目前人脸识别技术的应用非常广泛,虽然人脸数据的获取相当容易,但要想涵盖所有的应用场景,需要的数据量是巨大的。人脸跨域合成技术为数据有限的应用场景提供了一种数据增强的方法,即利用应用场景中原本少量人脸图像,合成大量人脸图像,增强数据的丰富度和多样性,提升人脸识别等系统的鲁棒性[24]。例如,基于前述的生成模型,可以从单一或少量表情、姿态、光照类型的人脸图像,合成出大量表情、姿态、光照变化的人脸图像,实现训练数据的大规模增长,提升模型的鲁棒性。类似地,由于人脸欺骗样本和对抗样本的存在,给人脸分析与识别系统的安全性带来挑战,但模型训练阶段人脸欺骗样本和对抗样本的类型是难以穷举的,因此可以利用跨域合成方法进行离线或在线的样本生成,使模型能更好地识别新的人脸欺骗和对抗样本类型。

4 人脸跨域合成技术发展趋势与挑战

人脸跨域合成虽然取得了飞速的进步,并在一些领域开展了应用,但随着更多更复杂应用场景的

出现,对人脸跨域合成的算法效率、合成质量的要求也越来越高。人脸跨域合成技术仍然面临诸多挑战,需要持续的研究。

1)自监督与弱监督跨域合成。现有的很多人脸跨域合成方法仍然依赖于来自不同域的成对的人脸图像训练,但在实际场景中往往缺乏配对的训练数据。例如在人脸画像—照片跨域合成、人脸跨行为表情合成等任务中,常见的数据集可能只包含1千人左右的图像,采用全监督学习的方法往往存在模型过拟合的问题,模型在新场景新数据上表现不佳。与此同时,非配对的人脸图像随处可见,可以轻而易举地获取上百万甚至上千万的非配对人脸图像。这些图像也包含了丰富的姿态、表情、行为变化,但只看一个个体可能包含的姿态、表情、行为变化非常有限,甚至只有1张照片。如何利用这些海量的人脸照片,提升人脸跨域合成的准确率和泛化性是未来值得研究的重要方向,研究基于自监督与弱监督学习的人脸跨域合成是充分利用海量非配对人脸图像的一种可能的方法。如何设计自监督与弱监督方法,从海量非配对人脸图像挖掘出有用的信息,仍然处于研究的起步阶段。

2)基于预训练大模型的跨域合成。大模型通常是基于大规模数据集进行训练的,模型中通常包含丰富的图像基础语义信息,这些基于语义信息推广到人脸图像领域可能也是适用的。例如:近期由Meta开源的图像分割大模型SAM(segment anything model)[25],实现了强大的零样本泛化能力,在自然图像、遥感图像等数据上都展现了良好性能。这样的大规模预训练模型,可以为人脸跨域合成任务提供丰富的底层语义一致性约束,提高跨域合成图像的质量。此外,利用预训练的文本—图像大模型[26],可以实现人脸图像更精细的语义描述,并将其与文本—图像大模型的语义空间对齐,提升人脸跨域合成的效果。通过融合预训练大模型的强大语义表达能力,有助于实现小样本甚至零样本下的高效准确的人脸跨域合成,拓宽人脸跨域合成的应用场景。但人脸图像包含的语义信息与自然图像相比,可能相对较少,如何在这种情况下利用好预训练大模型仍然是一个开放性问题。

3)基于跨域合成的隐私保护。随着人脸识别和分析技术的成熟,人脸图像成为日常工作生活中最广泛使用的人体特征数据,在安防、金融、交通、楼宇、零售、广告、教育、医疗、娱乐、政务等各个领域都在使用人脸比对、人脸认证、人脸画像等技术。因此,一个人的人脸图像可能被存储于不同企业(机构)的设备中,一旦泄露不光会引起个人肖像权问题,还可能带来身份被冒用滥用的风险。各国都越来越重视人脸图像等数据的保护问题,中国第十三届全国人民代表大会常务委员会第三十次会议通过《中华人民共和国个人信息保护法》,确立了专门的与人脸识别技术相关的个人信息保护规则和标准。为了保护个人信息不被无授权地冒用滥用,研究者开始关注人脸图像隐私保护问题,一种方式是基于对抗样本生成方法,为人脸图像加入肉眼难以分辨的微小扰动,使其他人脸识别系统失效[27],以及隐去非业务相关的人脸信息,但如何让通过隐私保护方法合成的人脸样本,对多种未知人脸识别系统都有效,仍然面临很大的挑战。

5　结论

人脸跨域合成涵盖的图像域丰富多样,由此衍生了不同类型的人脸跨域合成任务,推动了一系列人脸跨域合成技术的发展,带来了不同场景的应用。虽然人脸跨域合成技术飞速发展,但仍然存在标注数据和成对数据不足、算法效率和合成质量难以平衡、语义信息融合不足等问题,如果学术界和工业界能对上述问题给予更多的关注和研发投入,将有助于加速人脸跨域合成技术的成熟和落地,提升中国在人脸跨域合成等人工智能生成内容领域的国际竞争力。

参考文献(References)

[1] Blanz V, Vetter T. A morphable model for the synthesis of 3D faces[C]//Proceedings of the 26th annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 1999: 187-194.

[2] 苏从勇, 庄越挺, 黄丽, 等. 基于正交图像生成人脸模型的合成分析方法[J]. 浙江大学学报(工学版), 2005, 39(2): 175-179.

[3] Tran A T, Hassner T, Masi I, et al. Regressing robust and discriminative 3d morphable models with a very deep neural network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 1493-1502.

[4] Tang X O, Wang X G. Face photo recognition using sketch[C]//Proceedings of International Conference on Image Processing. Piscataway: IEEE Press, 2002.

[5] Williams I. Performance-driven facial animation[C]//SIGGRAPH '06: ACM SIGGRAPH 2006 Courses. New York: ACM, 2006.

[6] Gleicher M. Animation from observation[J]. ACM SIGGRAPH Computer Graphics, 1999, 33(4): 51-54.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Montreal: Curran Associates Inc., 2014: 2672-2680.

[8] Karras T, Laine S, Aila T M. A style-based generator architecture for generative adversarial networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 4396-4405.

[9] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[C]//Vedaldi A, Bischof H, Brox T, et al. European Conference on Computer Vision. Cham: Springer, 2020: 405-421.

[10] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [C]//Advances in Neural Information Processing Systems. Virtual: Curran Associates Inc., 2020: 6840-6851.

[11] Yu S K, Han H, Shan S G, et al. CMOS-GAN: Semi-supervised generative adversarial model for cross-modality face image synthesis[J]. IEEE Transactions on Image Processing, 2023, 32: 144-158.

[12] Hou A, Zhang Z, Sarkis M, et al. Towards high fidelity face relighting with realistic shadows[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 14714-14723.

[13] Wei Y X, Liu M, Wang H L, et al. Learning flow-based feature warping for face frontalization with illumination inconsistent supervision[C]//European Conference on Computer Vision. Cham: Springer, 2020: 558-574.

[14] Shen Y J, Yang C Y, Tang X O, et al. InterFaceGAN: Interpreting the disentangled face representation learned by GANs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2004-2018.

[15] Thies J, Zollhöfer M, Stamminger M, et al. Face2Face: Real-time face capture and reenactment of RGB videos [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2387-2395.

[16] Saharia C, Ho J, Chan W, et al. Image super-resolution via iterative refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 4713-4726.

[17] Yuan X W, Park I K. Face de-occlusion using 3D morphable model and generative adversarial network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2020: 10061-10070.

[18] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[DB/OL]. arXiv preprint: 1412.6572, 2014.

[19] Deng Y, Yang J L, Xu S C, et al. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 285-295.

[20] Tang X O, Wang X G. Face sketch synthesis and recognition[C]// Proceedings Ninth IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2008: 687-694.

[21] Kingma D P, Welling M. Auto-encoding variational bayes[C]// International Conference on Learning Representations 2014. Banff, AB, Canada: 2014.

[22] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[23] 肖冰. 人脸画像——照片的合成与识别方法研究[D]. 西安: 西安电子科技大学, 2010.

[24] 黄法秀, 张世杰, 吴志红, 等. 数据增广下的人脸识别研究[J]. 计算机技术与发展, 2020, 30(3): 67-72.

[25] Kirillov A, Mintun E, Ravi N, et al. Segment anything [DB/OL]. arXiv preprint: 2304.02643, 2023.

[26] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [DB/OL]. arXiv preprint: 2103.00020, 2022.

[27] 马玉琨. 基于人脸的安全身份认证关键技术研究[D]. 北京: 北京工业大学, 2018.

Research progress and trend of cross-domain face synthesis technology

LIU Qi[1], WU Haozhan[2], XIE Tianxin[2], HAN Hu[3]

1. Department of Network Security, Henan Police College, Zhengzhou 450000, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract Advances in deep learning technology and the development of the digital economy have promoted the development of artificial intelligence-generated content (AIGC) technologies such as virtual humans. Cross-domain face synthesis is one of the key technologies in virtual human production, and it has a wide range of applications in social media, film and television production and other fields. This paper summarizes the origin of cross-domain face synthesis technology, and its typical task types and difficulties, technological development and challenges, potential applications, and issues, and discusses its future development trend and challenges from the aspects of self-supervised and weakly supervised cross-domain synthesis, utilization of pre-trained large models, and privacy protection.

Keywords cross-domain face synthesis; virtual human; generative adversarial network; neural rendering; diffusion model

（责任编辑　王微）