中国科学: 信息科学

SCIENTIA SINICA Informationis

论文

基于纹理与几何解耦的说话人视频连续情感编辑模型

吕天[1], 温玉辉[2,3*], 孙志尧[1], 刘永进[1*]

1. 清华大学计算机科学与技术系, 北京 100084
2. 北京交通大学计算机与信息技术学院, 北京 100044
3. 交通数据分析与挖掘北京市重点实验室, 北京 100044
* 通信作者. E-mail: yhwen1@bjtu.edu.cn, liuyongjin@tsinghua.edu.cn

国家自然科学基金（批准号：62202257）和中国博士后科学基金（批准号：2021M701891）资助

摘要 说话人视频的情感编辑是计算机视觉和图形学当前研究热点之一, 其目的是将一段中性情感的人物说话视频转为带有目标情感的说话视频。已有的方法难以同时兼顾高清晰度情感编辑、人脸三维属性的保持以及模型适用不同的目标人物。为同时满足上述要求, 本文提出基于 Basel 人脸模型（BFM）条件的几何编辑网络作为几何情感编辑模块, 保证了几何编辑在不同目标人物场景下的通用性; 提出了基于人物分类器的纹理情感编辑模块, 使得精细纹理的编辑可以迁移到多人任务之中, 突破了以往情感编辑模型仅适用特定目标人物或适用多人模型生成质量不高的局限性。本论文提出的模型可以实现连续情感编辑强度的效果。实验结果表明, 本文提出的通用情感编辑模型在多人任务上的清晰度、人物保真度、情感编辑质量等各项指标均优于已有可适用于多人情感编辑的方法, 并且在训练集中未出现的目标人物上也能实现自然的情感编辑, 甚至在未见的人脸位姿的说话视频中也能达到合理的结果。

关键词 情感编辑、三维重建、深度学习、计算机视觉、神经网络

1 引言

近年来, 人脸情感编辑在计算机视觉与图形学研究领域引起了广泛关注, 其目的是根据目标情感编辑人物照片或者视频中的人脸属性（皱纹、表情细节等）, 生成符合目标情感的人脸照片或视频编辑结果。在虚拟主播和虚拟客服等实际应用中, 情感编辑模型能够为虚拟形象引入面部情感, 从而增强虚拟形象的真实感。针对人脸情感编辑任务, 本文提出基于 Basel 人脸模型（BFM）条件的几何

引用格式: 吕天, 温玉辉, 孙志尧, 等. 基于纹理与几何解耦的说话人视频连续情感编辑模型. 中国科学: 信息科学, 在审文章
Tian L, Yu-Hui W, Zhiyao S, et al.
(in Chinese). Sci Sin Inform, for review

编辑网络以及基于人物分类器的纹理情感模块, 对人脸几何和纹理进行解耦编辑。本文在满足高清晰度的情感编辑的基础上, 提出的模型适用多目标人物场景。

随着深度神经网络的发展, 近年来研究者提出了一些情感编辑模型。其中一部分是在图像层级直接利用神经网络编辑, 这种思路尽管能够达到较好的清晰度和纹理细节, 但是并未考虑人脸的三维属性, 导致人脸姿势变化时模型难以正确编辑 [6,10]。另一部分则是考虑利用三维重建来获取人脸几何以引入人脸的三维属性, 这些方法大多需要利用神经网络来表示特定人脸的精细纹理, 但是难以同时表示多个人物的精细纹理, 导致这些模型通常仅能适用于单个人物 [1,2]。部分工作尽管尝试解决精细纹理表示的问题, 但是却未能保证其中几何、纹理编辑能够适用于多目标人物场景中 [3]。由此可见, 已有的情感编辑模型很难同时满足图像编辑质量高、保持人脸三维属性、模型通用性等多个要求, 例如 [2] 等方法为满足特定人物的精细纹理而损失了情感编辑模型的通用性。然而, 虚拟主播等实际场景当中通常要求模型能够同时满足上述要求。因此, 构建一种能同时对多个人物的多角度说话视频进行情感编辑、且编辑结果质量高的模型, 是目前的一大难题。

为此, 本文设计了一种通用情感编辑框架。这种通用性体现在三个方面: 第一, 训练集已见人物中的通用性, 即保持对所有在训练过程中见过的人物有高清晰度、高质量的情感编辑效果; 第二, 训练集中未出现人物的通用性, 即模型能对训练集中没有出现的人物也能有合理的编辑结果; 第三, 对于训练集中未见角度的通用性, 即对训练集中未出现的人脸姿势进行情感编辑并达到不错的效果。

基于该目的, 本文提出了一种基于纹理与几何解耦的情感编辑框架, 以增强情感编辑的通用性。模型的输入为人脸图片, 首先利用三维人脸重建技术提取人脸三维几何、人脸纹理等信息。针对人脸三维几何, 本论文设计了基于 BFM 条件的人脸几何编辑网络, 实现了多目标人物场景下的人脸几何形状编辑; 针对人脸纹理信息, 本文使用纹理生成对抗网络 (t-GAN) 和编码器结合的方式, 将人脸纹理的编辑操作转化为在 t-GAN 隐空间中对输入图片的隐编码进行操作。为了保证隐编码编辑方向的正确性, 本文基于 [4] 设计了基于人物分类器的隐向量编辑方向的求解方法, 解决了不同人物的编辑方向不一致、进而导致无法使用单个编辑方向进行多个人物的隐向量编辑的问题。为了保证模型能够在不同场景下的多人任务均取得较好的纹理编辑效果, 本文采取了预训练结合微调的方式, 首先在 FFHQ 数据集上对 t-GAN 进行预训练以保证其生成纹理图片的质量, 此后在下游任务上再对其进行微调, 使其适用于具体的场景 (如同一组光照、场景的人物等)。本文主要贡献包括:

1. 提出了可以应用于多人的说话人视频情感强度连续编辑模型。
2. 提出了基于人物分类器的情感向量编辑求解方法, 解决了多人任务中的编辑方向求解困难与编辑结果不真实的问题。
3. 提出了基于 BFM 条件的人脸几何编辑网络, 使得基于生成对抗网络模型的人脸几何编辑方式可以应用到多人任务当中。

2 相关工作

2.1 三维人脸重建

三维人脸重建通常是基于某种形式的包含人脸信息的输入 (单角度 RGB 图像、RGB-D 图像等), 得到具有三维几何信息的人脸模型。为了实现对人脸的建模, [18] 于 1999 年提出了三维可形变人

脸模型（3DMM）, 首次提出将人脸几何与纹理进行参数化的思想。[13] 于 2009 年提出了 Basel 面部模型（BFM）, 它基于 3DMM 的思想提供了一组开源的人脸数据集。此后 [19]、[20] 等工作对人脸中的表情信息也应用了参数化的思想, 并得到了一组可以用于建模不同表情的表情基。

基于参数化人脸的思想, 三维人脸重建的目标就转化为基于输入信息预测人脸的 3DMM 系数, 而其优化目标则是减少通过这组参数重建出来的人脸在几何、纹理等方面与输入信息之间的差异。这种差异通常是在将模型预测得到的 3DMM 系数进行几何、纹理重建后, 利用重新渲染的方式将渲染后的人脸与原图的人脸的比较中体现的。一些工作, (例如 [22]), 利用传统的几何优化方法进行 3DMM 系数的预测; 而近年来相关领域的多数工作均由基于深度学习的方法来实现, 包括 [23]、[15]、[24]、[25] 等; 其中基于 FLAME 人脸模型 [20] 的 DECA 方法 [24] 在 NoW 基准 [23] 中得到了最佳重建结果。但是由于 FLAME 模型的几何相对 BFM 模型而言较为粗糙（FLAME 模型面片、顶点数目相对较少）, 因此本文将使用 [15] 提出的三维重建模型与 BFM 人脸数据的结合来实现三维人脸重建的功能。

2.2 语音驱动的说话人视频生成

语音驱动的说话人视频生成的任务是基于一段语音以及一张参考人物的图像或图像帧序列, 生成该人物说出这段语音的视频。一些方法 (例如 [11]、[26] 等) 提出了直接在图像层级、像素层级进行变换的方式完成人物说话的目标; 同时, [27] 等方法提出基于三维人脸重建技术, 通过对 3DMM 系数的操控, 实现人物的说话动作。

但是这些方法几乎都没有考虑人物的情感因素, 这导致了生成视频中的人物缺乏情感, 产生一定程度的不真实感。因为该任务本身的困难性, 即语音相关的面部表情变化与情感相关的面部表情变化通常而言是耦合的, 因此模型难以在视频生成时一次性地推理出合理的情感变化。为此, [28] 提出了基于 GAN 的方法来监督生成视频中的情感因素的方法, 保证生成视频具有多样的人脸情感。但是该方法未考虑人脸的三维本质, 因此在人脸转动时可能会出现瑕疵与扭曲。[29] 等人提出了输入额外的情感状态隐编码来指导模型生成不同情感的视频, 但是该方法只考虑了人脸三维几何的形变而不包括纹理细节的变化, 导致结果出现一定程度的不真实感。[2] 提出了带有情感的语音驱动人物说话视频的生成路线, 其主要思路是利用三维重建得到的人脸, 通过语音抽取的情感进行编辑变换, 利用变换后的人脸几何提取边缘得到边缘图, 对边缘图利用 vid2vid [30] 的方法进行复原, 但是该方法需要针对每一个特定人物进行训练, 无法适用于多目标人物场景。

综上所述, 由于基于语音与参考帧直接生成带有情感的说话人视频相对比较困难, 因此本文提出将带有情感的语音驱动的说话人视频生成任务以分步的形式完成, 即首先生成中性说话视频, 再通过情感编辑生成带有情感的说话视频。本文主要针对第二步情感编辑开展研究工作。

2.3 人脸情感属性编辑

人脸情感编辑属于情感计算领域。情感计算目的是基于算法与模型, 使计算机具有识别、理解并表达人类情感的能力。情感计算领域包括人脸表情情感识别与理解 [5] [7]、脑电波情感识别与理解 [9] [8]、人脸表情生成 [1] [3] 等领域。其中本文研究的人脸情感编辑属于计算机表达人类情感方面, 通过情感编辑可对人物形象赋予多样多强度的面部表情。该任务可视为图像到图像的变换任务。其中的主要目标是在保证人脸身份信息、位姿信息等基本内容不变的条件下, 改变人脸的表情属性。

从编辑模态的角度来看人脸情感编辑包括图像层级以及视频层级的编辑。针对图像层级的人脸情感属性编辑任务, 一些基于 GAN 网络的方法被提出。例如 ExprGAN [6] 提出了表情控制模块将情感转为连续特征, 实现了人脸的多种强度的情感编辑效果, 但是生成人脸的清晰度不足; StyleRig [34] 提出了基于重建的三维人脸几何, 利用 StyleGAN 在隐空间编辑 3DMM 系数, 使其变换为目标情感的 3DMM 系数, 但是该方法并未考虑纹理信息, 导致生成的结果真实性下降; StarGAN [10] 提出了将不同域信息和图片一起输入进行训练, 并在域标签中加入掩模向量, 便于不同的训练集进行联合训练以达到图像翻译的效果, 并展示了部分情感编辑的结果; 该方法需要情感编码向量与人脸图片同时输入生成器来控制最终的情感编辑结果; 但是该方法未考虑人脸的三维属性, 因此在训练集人脸姿势有限的情况下泛化性较差。[35] 提出了将纹理与几何进行解耦并分别进行属性编辑的方法, 并可以应用于多人任务, 但是该方法并不能保证情感属性的连续编辑, 同时图像生成质量不高。为改进生成图像质量以及情感编辑的效果, 近年来研究者提出了 StyleGAN 以及其隐空间编码器 [16] [33] [17]。StyleGAN是一种图像生成的模型, 可以将隐空间中的向量解码为图片; 基于训练充分的 StyleGAN 生成器, 可训练一个编码器将输入图像映射到隐空间中合适位置, 使得生成器对该位置的隐编码的解码结果与输入图像保持一致 [17]。通过将输入图片映射到 StyleGAN 隐空间得到图像的隐编码, 并对这个隐编码进行特定方向的移动即可改变图像的特定属性。最终利用生成器解码即可得到经过属性编辑的图片。基于 StyleGAN 架构, 一些工作专注于研究如何优化将图像映射到隐空间的效果, 使得隐空间更加接近数据的特征空间, 例如 [31]、[32] 等; 而另一些工作则尝试基于编码器的设计, 将输入图片正确地编码到 StyleGAN 隐空间当中, 例如 [33]、[17] 等。虽然 StyleGAN 可以生成高质量的图像, 但是其隐空间的属性编辑仍然是一个困难的问题, 因为其中仍然存在多种属性耦合的情况, 导致编辑目标属性时会影响到原图的其他属性, 特别是在生成对象具有较强多样性的情况下。上述方法中, StarGAN、ExprGAN、基于 StyleGAN 的编辑方法均可实现连续情感的编辑。其中 StarGAN、ExprGAN 通过可以连续变化的情感控制编码与人脸图像共同输入到生成器中, 实现连续受控的情感编辑; 而基于 StyleGAN 的编辑方法则可通过连续变化图像隐编码沿编辑方向的移动距离, 实现连续程度的情感编辑。

针对视频层级的情感编辑任务, [1] 提出了动态神经纹理的思想, 通过三维重建得到的人脸几何标记面部区域, 利用神经纹理以及表情标签在标记区域生成目标情感的人脸; [3] 提出的方法将 StyleGAN 应用于人脸原图纹理的编辑, 并使用 StarGAN 的方式对人脸的 3DMM 系数进行情感编辑, 并利用平滑算子保证帧间连续性。然而上述方法针对视频的情感编辑方法只能应用于单个人物, 不能作为通用模型应用于多人任务上。

3 方法

为实现多人连续情感编辑的效果, 本文提出了图1所示的模型架构。该模型输入为红色框对应的两部分, 即一个带有中性情感的人物说话视频以及一个目标情感。首先模型将目标情感编码为向量, 同时三维重建模块对视频的每一帧进行三维重建得到每一帧人脸的 BFM 系数以及原图纹理。对于 BFM 系数, 需要将其输入到基于 BFM 条件的几何编辑网络中得到经过情感编辑的人脸几何; 对于原图纹理, 首先输入到纹理编码器中得到人脸纹理的 t-GAN 隐空间编码, 并利用情感编码驱动隐空间编码沿着指定方向移动完成纹理编辑, 最后使用 t-GAN 生成器对编辑后的隐编码进行解码, 得到情感编辑后的纹理贴图。本文提出的模型先利用平滑化模块来平滑编辑后的 BFM 系数以及纹理贴图, 再使用可微渲染器结合平滑后的 BFM 系数与纹理得到初步渲染结果, 输入到牙齿补全模块中进

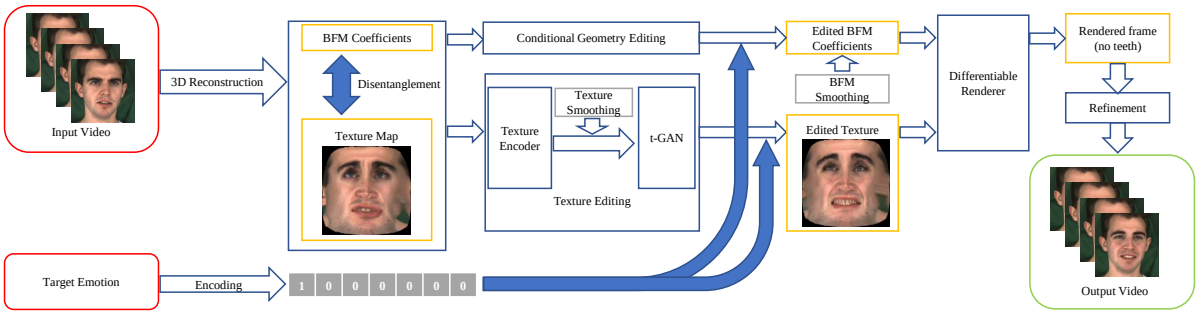


图 1　本文提出的说话人情感编辑模型框架

Figure 1　The proposed framework of talking head emotional editing

行牙齿的补全, 最终得到输出的情感编辑视频。

三维人脸重建　本文参考目前已有的 Deep Face Reconstruction 这一工作的实现三维人脸重建 [15], 该工作提出的模型简单有效, 并且可以利用大量无标注的人脸图像进行自监督训练, 具有较强的泛化性以及重建精度。使用了 300WLP [37]、celebA-HQ [38]、FFHQ [16]、LFW [39]、LS3DW [40]、UTK [41] 等数据集完成三维人脸重建模块的训练。此后的实验中该模型的参数固定, 不再改变。为满足三维人脸重建模型对于输入图像对齐的要求, 因此在训练前的数据预处理过程中应当先对所有输入帧进行人脸对齐, 后经裁剪与变换后得到用于后续处理的人脸照片。本文使用的 BFM 模型提供了一种平均人脸几何 $\bar{\mathbf{S}}$, 包含其中顶点的空间坐标和连接关系, 以及人脸身份几何 PCA 基 $\mathbf{B}_{\mathrm{id}}$ [13] 和人脸表情几何 PCA 基 $\mathbf{B}_{\mathrm{exp}}$ [19]。本文采用的三维人脸重建方法可以得到一组人脸 BFM 系数 [13], 包含人脸的形状系数 α 和表情系数 β, 可以重建出特定 BFM 系数下人脸的几何模型, 如公式1所示。

$$\mathbf{S} = \bar{\mathbf{S}} + \mathbf{B}_{\mathrm{id}}\alpha + \mathbf{B}_{\mathrm{exp}}\beta. \tag{1}$$

纹理贴图提取　由于 BFM 系数生成的人脸纹理较为模糊、缺失高频信息, 因此本文使用原图纹理提取的方法生成纹理贴图。首先将位于相机坐标系中的三维人脸投影到图像空间中, 并利用人脸网格中的顶点在图像空间中的坐标, 在原图中进行采样, 得到原图纹理。由于 BFM 人脸数据集中的所有人脸结构相同, 因此本文采用一个预定义贴图模板, 使得不同人脸的相同序号的顶点的纹理坐标保持一致, 从而得到格式一致的纹理贴图。基于上述原图采样的纹理, 以及顶点之间的连接关系, 使用重心插值的方式通过顶点颜色获得完整的原图纹理。图1中展示了纹理贴图提取的结果。

情感编码模块　本模块对情感进行向量化编码。设当前人物情感类别序号为 $i, i = 0, 1, \ldots, E_t - 1$, 情感强度为 $j, 0 \leqslant j \leqslant E_m$, E_t 为除中性情感以外的情感类别总数、E_m 为情感强度的最大值, 其中情感类别总数 E_t 来源于使用的 MEAD 数据集, 包含 7 种（开心、生气、悲伤、惊讶、害怕、恶心、轻蔑）能够覆盖日常生活中绝大多数应用场景的情感表达 [14]。情感强度最大值 E_m 定义与 MEAD 数据集定义一致, 如章节4.1所示。最终本模块将输入情感编码为 $(e_0, e_1, e_2, \cdots, e_{E_t-1})$。

$$e_k = \begin{cases} 0, & i \neq k. \\ \frac{j}{E_m}, & i = k. \end{cases} \tag{2}$$

3.1 基于人物分类器的纹理贴图编辑模块

为了实现从中性人脸纹理到目标情感人脸纹理的编辑, 本文首先预训练一个纹理生成对抗网络（t-GAN）, 用于生成高清、种类广泛、细节丰富且包含情感成分的人脸纹理。而情感编辑的操作, 则是在 t-GAN 隐空间中沿特定方向移动中性人脸纹理的隐编码, 使得移动后的隐编码经 t-GAN 解码后生成带有情感的人脸纹理, 并保持人物相关的属性不变。为了利用 t-GAN 的生成能力, 本文需要设计一个编码器将人脸纹理映射到隐空间中的正确位置, 使得解码器对该位置的编码解码后能够得到近似原来的人脸纹理。本文中的人脸纹理属于细节丰富的图片, 需要生成器具有较强的解码能力; 鉴于 StyleGAN 能够生成高清的图片, 并且其隐空间中特征属性解耦程度较好, 因此本文参考该工作实现 t-GAN [16]。为保证编码器的性能以及映射能力, 本文参考 [17] 的工作设计 t-GAN 的编码器。待上述模型收敛后, 编码器可以有效地将不同情感强度的人脸纹理映射到 t-GAN 隐空间中的正确位置, 基于此可进行后续情感编辑方向的计算。

由于 t-GAN 隐空间中, 以特定人物作为条件的情感纹理分布可能有所差异, 这导致不同人物的情感编辑方向存在差异。为解决该问题, 本文引入一个基于纹理贴图的辅助人物分类器, 用于判别当前贴图的人物。设数据集有 N 个人物, 在 t-GAN 隐空间中, 从该人物 i 某一张中性情感纹理的隐编码 $\mathbf{x}_{0i}$ 到某一情感 j 的纹理的编辑向量为 $\mathbf{d}_{ji} \in \mathbb{R}^{16\times512}$, 则纹理情感编辑的结果为:

$$\mathbf{x}_{ji} = \mathbf{x}_{0i} + \alpha\mathbf{d}_{ji}, \tag{3}$$

其中 α 为用户预先指定的情感编辑的强度。尽管可对数据集中已见人物分别求解其各个情感的编辑方向, 达到最为理想的编辑效果, 但是如此无法对未见人物进行情感编辑。为此本文假设: 中性人脸纹理类似的人物, 在情感变化时也有类似的变化; 同时仿照 3DMM, 人脸纹理可由有限个模板纹理进行加权求和近似表示。由此可得到通用情感编辑的方法: 给定一张人脸纹理 I, 输入此前的分类器 f 中。f 在给出分类结果前需要得到一个经过 Softmax 归一化操作的向量, 将其中各个分量作为对应模板纹理的权重系数, 则 I 的隐编码 $\mathbf{x}_0$ 对于第 j 种情感的的情感编辑方式为

$$\mathbf{x}_j = \mathbf{x}_0 + \sum_{v=1}^{N} \mathrm{Softmax}(f(I))_{v-1}\mathbf{d}_{jv}, \tag{4}$$

其中, v 为迭代变量, 代表训练集中第 v 号人物, $1 \leqslant v \leqslant N$, N=39 。对于特定人物 i, 其所有中性人脸纹理贴图在 t-GAN 隐空间中的坐标分别为 $(\mathbf{x}_{0i1}, \mathbf{x}_{0i2}, \cdots, \mathbf{x}_{0iN})$, 其中 $\mathbf{x}_{0ir} \in \mathbb{R}^{16\times512}, 1 \leqslant r \leqslant N$, 其第 j 种情感在隐空间中为 $(\mathbf{x}_{ji1}, \mathbf{x}_{ji2}, \cdots, \mathbf{x}_{jiM}), \mathbf{x}_{jir} \in \mathbb{R}^{16\times512}, 1 \leqslant r \leqslant M$。分别计算第 j 号情感, 且情感强度最大的所有纹理贴图的隐空间坐标的平均值, 同时计算中性情感对应的纹理贴图的隐空间坐标平均值, 作差后即可得到第 i 个人物的由中性情感指向第 j 号情感的编辑向量。

$$\mathbf{d}_{ji} = \frac{1}{M}\sum_{r=1}^{M}\mathbf{x}_{jir} - \frac{1}{N}\sum_{r=1}^{N}\mathbf{x}_{0ir}. \tag{5}$$

相比于近年来基于 StyleGAN 的方法 [17], 本文利用了纹理贴图的结构一致性降低了无关特征属性（背景、人脸姿势、头发等）的干扰, 并且本文使用了基于人物分类器的情感编辑方法, 使得模型对于各个人物的情感编辑效果更为准确; 相比于 StarGAN [10], 本文方法无需对生成器输入情感控制编码, 而是在隐空间中依据目标情感及其强度连续变化隐编码。

3.2 人脸几何编辑模块

基于 BFM 条件的人脸几何编辑网络 该模块基于人脸身份、表情系数、目标情感, 输出目标情感下人脸的身份、表情系数。考虑到人物不同情感下语速的差异, 因此难以获得成对训练数据。为此本文参考 [12] [10] , 实现了一种无需成对数据的基于 BFM 条件的几何编辑网络。其中引入条件是防止生成器将 BFM 系数映射到其他人物目标情感的 BFM 系数上, 同时使判别器对 BFM 系数真假做出更准确的判断。对于判别器 D, 条件 BFM 系数的选取规则为：接受真输入时, 采取输入 BFM 系数对应人物的一个随机中性 BFM 系数；接受假输入, 采取生成器使用的条件 BFM 系数。而生成器 G 的条件 BFM 系数采取输入 BFM 系数对应的人物的随机一个中性 BFM 系数。

判别器损失函数 设判别模型的待检验 BFM 系数为 s, 并且真输入 BFM 系数为 s_t, 对应情感为 e_t, 条件 BFM 系数为 s_{rt}；假输入 BFM 系数为 $s_f = G(s_t, \hat{e}, s_{rf})$, s_{rf} 为条件 BFM 系数, $\hat{e}$ 为目标情感向量。判别器与生成器分别记作 D, G, 而 D_c 代表判别器 D 对输入图像的真假判别结果, D_e 代表判别器 D 对输入图像的情感分类结果。首先, 情感分类损失用于衡量模型对 BFM 系数进行情感预测的能力, 目的是增强模型的判别力；该损失只有真输入时才会计算。

$$L_{\text{emo}}^{D} = \|D_e(s_t, s_{rt}) - e_t\|_2^2. \tag{6}$$

判别损失, 其实现参考 [10]。

$$L_{\text{dis}}^{D} = -D_c(s_t, s_{rt}) + D_c(s_f, s_{rf}). \tag{7}$$

梯度惩罚项, 其目的是使模型收敛更加稳定。其中将真假输入进行叠加 $\tilde{s} = \alpha s_t + (1-\alpha)s_f$, 将判别器对其判断结果进行求导, 表示判别器对扰动的敏感性 [46]。该项具体如下所示。

$$L_{\text{gp}}^{D} = \left(\left\| \frac{\partial D(\tilde{s}, s_{rt})_c}{\partial \tilde{s}} \right\|_2 - 1 \right)^2. \tag{8}$$

最终用于训练判别器的损失函数为如下所示, 其中 λ_e, λ_{gp} 为权重超参数。

$$L^D = L_{\text{dis}}^{D} + \lambda_e L_{\text{emo}}^{D} + \lambda_{\text{gp}} L_{\text{gp}}^{D}. \tag{9}$$

生成器损失函数 设输入给生成器的目标情感为 $\hat{e}$, 原始情感为 e_o, 其余符号定义与判别器相同。首先参考 [12] 的设计, 给出如下判别器反馈损失项：

$$L_d^G = -D_c(G(s_t, \hat{e}, s_{rt})). \tag{10}$$

生成器的情感分类损失项, 用于鉴别生成器生成系数对应的情感是否符合预期：

$$L_{\text{emo}}^{G} = \|D_e(G(s_t, e, s_{rt})) - \hat{e}\|_2^2. \tag{11}$$

循环重构损失项, 是利用生成器生成目标情感的 BFM 系数, 再用生成器基于原始情感重构, 通过比较重构前后人脸几何顶点的空间坐标变化衡量损失。其中 recon(·) 代表用 BFM 系数获取几何。

$$L_g^G = \|\text{recon}(G(G(s_t, \hat{e}), e_o)) - \text{recon}(s_t)\|_2^2. \tag{12}$$

区域关键点损失项, 通过赋予面部不同区域不同权重, 使生成器关注对情感编辑结果产生较大影响的区域（如眼部等）。损失函数表达式中 K_q 代表第 q 个区域的关键点集合, λ_j 为第 q 个区域的权重, 而 $p_k, \hat{p}_k$ 则代表重建几何中的第 q 个区域的第 k 号关键点坐标的重构值与真实值。

$$L_{\mathrm{id}}^{G} = \sum_{q} \sum_{k \in K_q} \frac{\lambda_q}{|K_q|} \|p_k - \hat{p}_k\|_2^2. \tag{13}$$

唇形保持损失项, 用于防止情感编辑过度破坏唇形, 以保证唇形与说话内容的一致性。本文参考 [21] [3] 来约束嘴唇的开合程度来保持这种一致性。此处本文分别选取第 52 与 58 号、第 63 与 67 号关键点组成的点对, 分别表示嘴唇外部位于中轴线上的两点, 以及嘴唇内部位于中轴线上的两点, 它们能较好地衡量嘴唇的开合程度, 具体说明如补充材料 1.3.2 节所示。基于此分别计算真实 BFM 系数与重构 BFM 系数的两对关键点的距离, 比较各自距离是否发生显著变化, 来衡量唇形的保持状况, 即:

$$L_{\mathrm{lip}}^{G} = \| \|p_{52} - p_{58}\|_2^2 - \|\hat{p}_{52} - \hat{p}_{58}\|_2^2 \|_1 + \| \|p_{63} - p_{67}\|_2^2 - \|\hat{p}_{63} - \hat{p}_{67}\|_2^2 \|_1. \tag{14}$$

正则化损失项, 用于保证情感编辑后的人脸几何的身份信息, 不产生异常值导致形状发生剧变.

$$L_{\mathrm{reg}}^{G} = \|\mathrm{recon}(G(s_t, e)) - \mathrm{recon}(s_t)\|_2^2. \tag{15}$$

最终用于生成器的损失函数如下所示, 其中 $\lambda_{\mathrm{lip}}, \lambda_{\mathrm{reg}}, \lambda_e, \lambda_g$ 为各损失项的权重系数:

$$L^{G} = L_d^{G} + \lambda_e L_{\mathrm{emo}}^{G} + \lambda_g L_g^{G} + L_{\mathrm{id}}^{G} + \lambda_{\mathrm{lip}} L_{\mathrm{lip}}^{G} + \lambda_{\mathrm{reg}} L_{\mathrm{reg}}^{G}. \tag{16}$$

3.3 细化模块

牙齿补全模块 该模块用于修补由于 BFM 模型未对牙齿部分建模而产生的嘴部空洞。首先, 在数据预处理部分中, 使用 BFM 模型 [13] 中包含的人脸 68 个关键点中的嘴部关键点在世界坐标系下的坐标, 并且结合三维人脸重建模型得到的相机参数计算嘴部关键点在图像空间中的坐标。获取嘴部关键点图像坐标后, 可仿照人脸对齐的方式 [15] 对嘴部进行对齐, 后经过裁剪与变换即可得到用于该模块的训练数据, 包括完整的嘴部原图以及去除牙齿区域的待补全图片组成的训练对。基于该训练对, 可通过预训练 t-GAN 结合编码器的方法, 将无牙齿图像映射到 t-GAN 隐空间中, 并且经过 t-GAN 解码后得到有牙齿的图像, 从而实现将无牙齿嘴部进行补全。

平滑化模块 该模块通过对视频帧经过情感编辑得到的 BFM 系数以及纹理的 t-GAN 隐编码, 分别进行相邻帧的平滑化处理, 保证输出视频的帧间连续性。其中在编码空间进行平滑化, 可以避免在像素空间中进行平滑化而产生的运动模糊（眼动等）。假设目前处理帧序号为 g, 平滑窗口的单边大小为 l, 则窗口大小为 $2l+1$。设第 g 帧情感编辑的 BFM 系数为 (α_g, β_g), t-GAN 隐空间编码为 $\mathbf{x}_g$。对于 BFM 系数, 本文仅平滑其中的表情系数而不平滑身份系数, 这是为了防止因过度平滑导致唇形、口形失真的情况。本文设计的平滑方式, 形式化表示如下:

$$\tilde{\beta}_g = \sum_{y=g-l}^{g+l} w_y \beta_y, \qquad \tilde{\mathbf{x}}_g = \sum_{y=g-l}^{g+l} w_y \mathbf{x}_y, \tag{17}$$

其中, 初始权重系数的求解方式采取余弦权重的方式, 模拟帧贡献度按照中间高两边低的分布情况, 具体如下所示:

$$w'_y = \frac{1}{2}\left(1 - \cos\left(\frac{\pi(y-(g-l-1))}{(l+1)}\right)\right), g-l \leqslant y \leqslant g+l \tag{18}$$

, 最终的权重系数需要进行归一化, 防止出现平滑异常:

$$w_y = \frac{w'_y}{\sum_z w'_z}, g-l \leqslant z \leqslant g+l. \tag{19}$$

4 实验

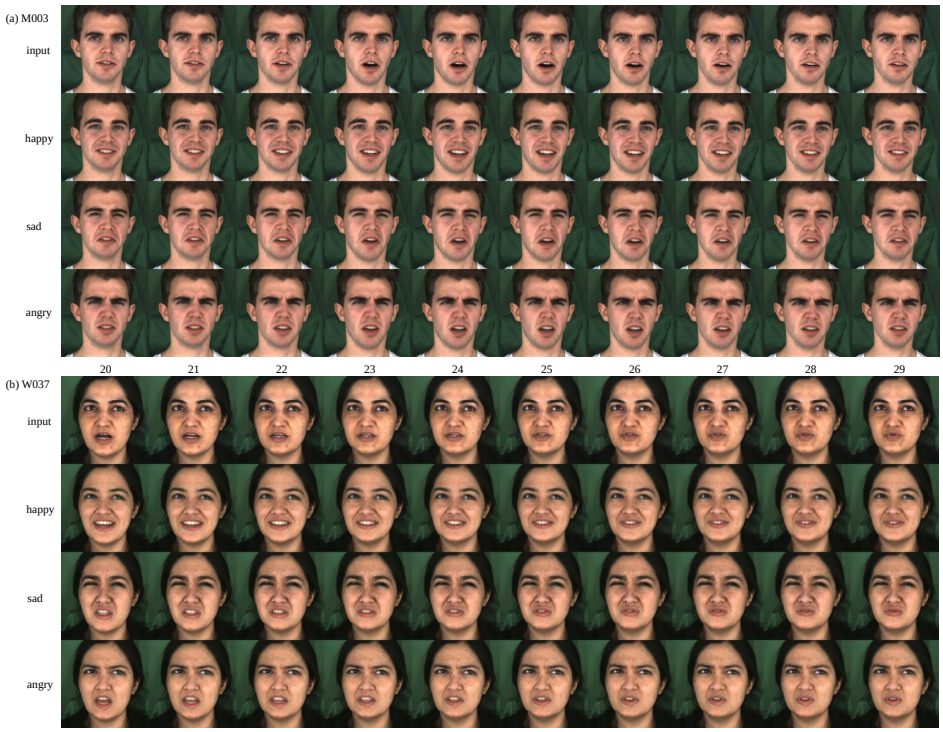


图 2 模型视频连续帧情感编辑效果示例, 其中 (a) 为训练集中出现人物 M003 的示例, (b) 为训练集中未出现人物 W037 的示例。列表示视频帧的序号, 行表示目标情感的类别

Figure 2 Examples of emotional editing on continuous frames of talking videos. (a) represents examples of subject M003 which appears in training set, (b) represents examples of subject W037 which does not appear in training set. Columns stand for the index of frames and rows stand for the types of target emotion

训练流程 本论文的训练流程为: 训练三维人脸重建模型; 对 FFHQ、MEAD 数据集进行数据预处理; 使用 FFHQ 数据集对应的纹理贴图预训练 t-GAN; 使用 MEAD 数据集微调前一步对应的 t-GAN, 并其对应的纹理编码器; 训练牙齿补全的 t-GAN, 以及其对应的纹理编码器; 训练人脸几何编辑模块; 训练人脸纹理贴图的分类器; 最后分别求解每个人物的每种情感编辑方向。更多实验相关的细节与结果总结在本文的补充材料和演示视频中。

4.1 数据集

FFHQ FFHQ 数据集是一个高质量的人脸数据集, 包含 1024×1024 分辨率的 70,000 张高清人脸图像, 用于本论文的 t-GAN 预训练步骤。其中, 为保证模型生成正常的纹理贴图, 需要将有遮挡、光

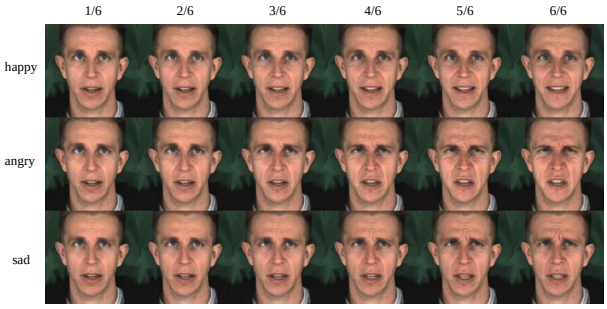


图 3 M007 人物不同强度、不同情感编辑结果, 其中列表示不同情感强度、行表示不同目标情感种类

Figure 3 The emotional editing results of subject M007 with different levels of intensity and different kinds of emotion. The columns represent different levels of intensity and the rows represent different kinds of target emotion

线异常、大角度姿势的图片进行剔除。最终本文使用 49612 张图片进行预训练。

MEAD 本文选取了 MEAD 数据集作为下游具体情感编辑任务的数据集 [14]。MEAD 定义了情感的强度, 表示数据集中的人物在说话时拥有某种情感的程度, 例如稍微高兴、非常高兴等。情感强度可在人类能分辨的基础上分为三个类别, 分别为弱、中、强。其中弱的情感强度下, 人脸会有细微而容易被察觉的面部表情变化; 中情感强度代表了人在该情感状态中的正常表现, 数据集中记录的该情感强度的视频可以作为这种情感的典型面部表现; 而强情感程度则描述了处于某种情感下夸张、剧烈的人脸表情, 通常对应非常明显的面部变化。因此, 最大情感强度对应的就是 MEAD 数据集中使用的强情感程度。经过数据清洗, 本文选取其中 47 个人物进行实验, 其中 39 个人物为作为微调的训练数据（已见人物）, 剩余 8 个为未见人物。每个人物均有 8 种不同情感的说话视频, 分别为中性、开心、生气、悲伤、惊讶、害怕、恶心、轻蔑, 除中性外的情感均有三种不同强度。其中训练过程仅采用正脸视角的数据, 在测试步骤中会使用其他视角的数据验证模型的通用性。

4.2 模型定性结果

已见人物与未见人物 本文展示已见人物 M003、未见人物 W037 在 0 号视频的第 20 至第 29 帧的 10 帧图片上的时序编辑结果, 分别如图2a、2b 所示。其中第一行为输入的原始中性情感视频帧、第二、三、四行分别为开心、悲伤与生气的情感编辑结果（均采取最大强度情感编辑）, 每一列从左至右对应按时间顺序排列的一帧。可见 M003 的情感编辑结果比较真实, 例如眼睛形状、嘴唇形状、皱纹等部分均符合目标情感。同时细节保留十分充分, 例如面部斑点、胡子等, 在编辑结果中十分清晰; 对于未见人物 W037, 模型仍能实现合理的编辑结果。人脸身份信息、面部细节在情感编辑的过程中也得到了较好的保留, 并且面部情感也与目标情感保持一致。

连续可控 本文模型可以连续地编辑人脸情感, 此处选取 M007 人物进行效果展示, 如图3所示。其中每一行表示不同情感, 从左至右第 $i, 1 \leqslant i \leqslant 6$ 列代表情感强度为 $\frac{i}{M}$, 其中 M 为实验中设定的最大情感强度。可见每一行从左至右情感强度的连续变化, 包括训练集中不存在的情感强度（奇数列）。

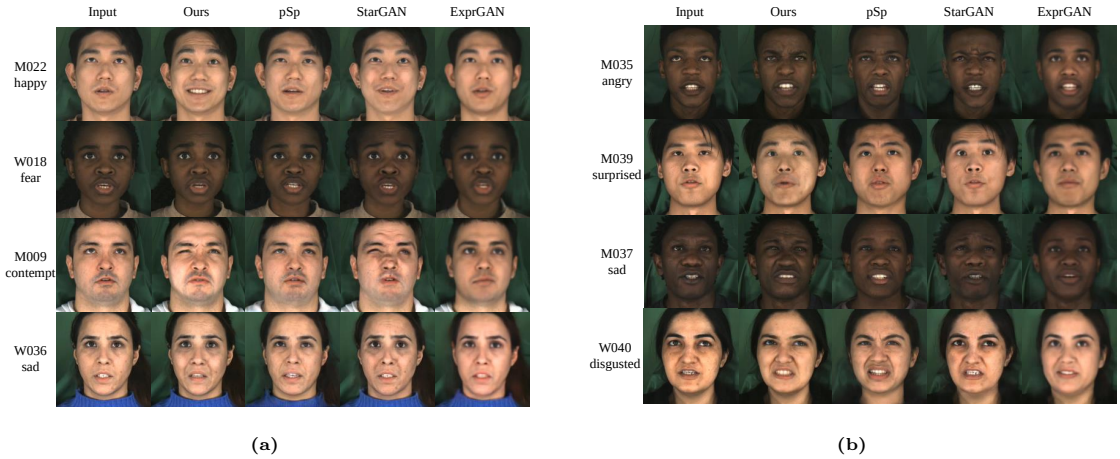


图 4　各方法对 (a) 已见人物和 (b) 未见人物编辑结果

Figure 4　The emotional editing results of different methods on (a) subjects appeared in training step and (b) subjects not appeared in training step

4.3　定量与定性对比实验

本文的对比实验主要在多人场景进行, 因此选取了StarGAN, ExprGAN 以及 pSp 作为本文的对比方法 [10] [6] [17]；而 EVP、DNT 等方法 [2] [1] 明确指出需要对单独人物进行训练, 将精细纹理存储在神经网络中, 因此直接在多人场景比较会产生不公平。

评测指标　本论文采取 FID、FED、CPBD、ArcSim 对情感编辑效果进行评测。FID 为 Fréchet Inception Distance, 是真实图像和生成图像的特征向量之间距离的度量, 用于评估生成图像的质量 [42]；FED 与 FID 类似, 差异在于 FED 使用的特征提取器是情感分类网络的编码层, 因此用于衡量生成情感的质量 [43]；CPBD 即模糊检测的累积概率, 用于衡量图像的清晰度 [44]；ArcSim 是通过人脸深度编码的比较来衡量人脸相似性的指标 [36]。

指标结果　本文对 MEAD 训练集中的所有 39 个人物, 使用不同模型生成其每一种情感以最大强度表现的说话视频, 每一个模型分别生成 273 个视频, 并对这些视频进行指标的计算, 结果如表1所示。未使用未见人物进行计算, 是考虑到未见人物的编辑结果即使合理, 但也可能与事实差异较大, 因此主要以可视化结果展示未见任务编辑效果。可见本论文提出的方法在各项评测指标均达到了最优的结果。

4.4　消融实验

已见人物、未见人物的可视化结果　图4展示了本文方法与其他情感编辑方法对已见人物（左）、未见人物（右）情感编辑的可视化结果对比。可以看出, ExprGAN 对两类人物编辑得到的图像较为模糊, 且情感编辑效果微弱, 出现了虚影、脸部变形等情况；pSp 编辑结果相对比较清晰, 但是情感编辑效果微弱, 并且对未见人物编辑结果不合理, 其中人物脸型、发型如输入图片不一致。而本文提出的方法和 StarGAN 均实现符合预期的情感编辑结果。但对于已见人物, StarGAN 在 M009 的编辑结

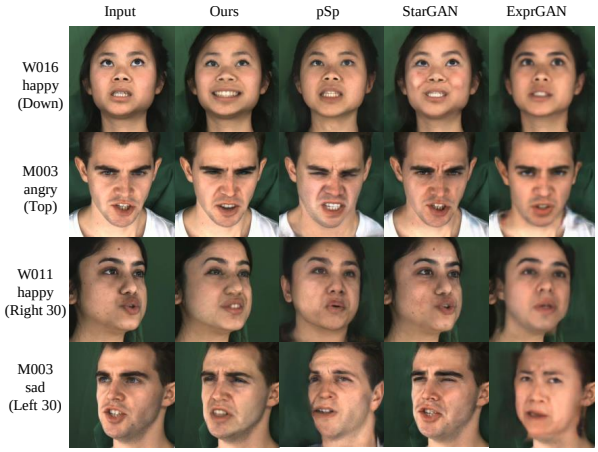


图 5　各方法对于未见位姿人脸的情感编辑结果

Figure 5　The emotional editing results of different methods on unseen poses of head

表 1　各模型应用于已见人物的情感编辑评测指标结果

Table 1　The evaluation results of emotional editing of different models on subjects appeared in training step

	Ours	StarGAN	ExprGAN	pSp
FED	**17.54**	21.84	33.47	29.56
FID	**87.35**	96.67	134.3	101.60
CPBD	**0.2316**	0.1748	0.1750	0.2286
ArcSim	**0.8548**	0.7828	0.7163	0.7549

果中出现了眼部、鼻梁上的严重变形, 并且 StarGAN 编辑结果清晰度相对较差, 导致 M022、W036 眼睛等部位的细节出现损失；对于未见人物 M035、M037, StarGAN 编辑结果丢失了眼部的细节出现模糊, 对于 W040, StarGAN 编辑幅度很弱, 例如眼部相比输入基本没有变化。而本文提出的方法仅在未见人物 M039 的编辑结果上有略微偏色, 但表情符合目标情感。由此可见, 本文提出的方法在相比基线能实现清晰度高、合理真实、情感强度符合要求的情感编辑效果。

训练集中未出现的人脸位姿　本文的方法实现了纹理与几何的解耦, 因此尽管本论文的模型是在正面视角的人脸数据上训练, 但模型具有在未见位姿上进行编辑的能力。图5展示了四种情感编辑方法对未见位姿人脸的情感编辑结果。可见 ExprGAN 在人脸位姿变化后无法进行正常的编辑；pSp 同样会出现人脸姿势、人脸形状的严重变形；而 StarGAN 在 W016 人物中出现了异常的面颊样式, 在 M003(Top) 以及 W011 的编辑结果中情感编辑幅度相比正面人脸显著减小。本文提出的方法则对于展示的四种新位姿, 编辑效果符合目标情感, 且细节表情真实, 没有出现明显的面部形变与虚影。

不使用分类器　取消分类器模块后, 需要使用训练集中所有人的数据求解一组统一的情感编辑方向。基于这种方式得到的编辑向量进行情感编辑, 编辑结果如图6所示。其中第一行、第二行分别为使用、不使用分类器进行编辑的结果, 每列表示不同情感。可以看出取消分类器后, 人物眼部出现黑色的晕

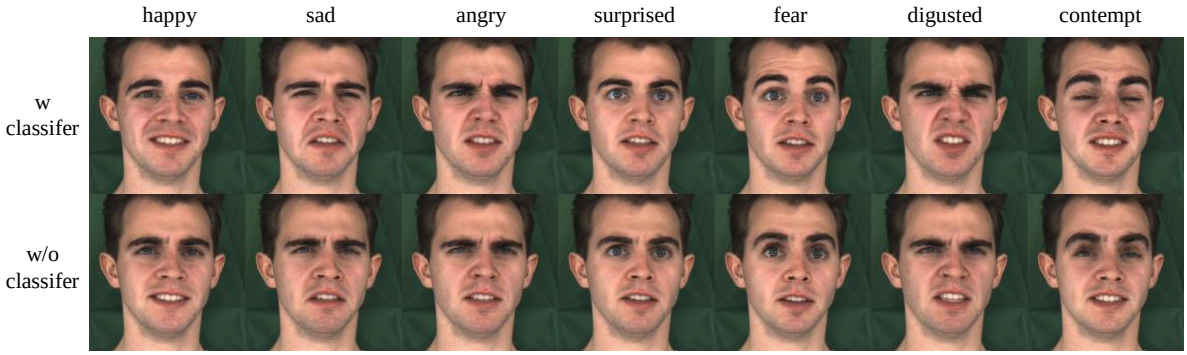


图 6　M003 在有分类器、无分类器情况下的情感编辑效果, 每一列表示不同的情感种类

Figure 6　The emotional editing results with and without classifier for subject M003. Columns represent different types of emotion

表 2　消融实验的评测结果, (a) 为有、无分类器的评测结果, (b) 为同时编辑纹理与几何、只编辑纹理、只编辑几何的评测结果

Table 2　The evaluation results of ablation study. (a) The evaluation results with and without classifier. (b) The evaluation results of editing texture and shape, editing texture only and editing shape only.

(a)

	w classifier	w/o classifier
FED	**17.54**	20.85
FID	**87.35**	89.97
ArcSim	**0.8548**	0.7980

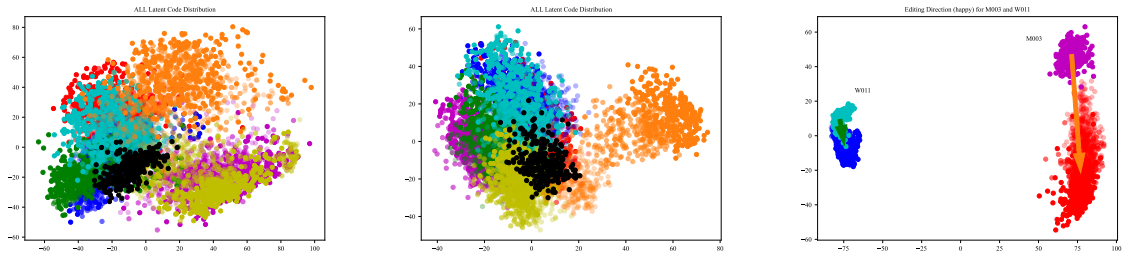
(b)

	All	w/o shape	w/o texture
FED	**17.54**	22.86	26.73
FID	**87.35**	88.90	89.39
ArcSim	**0.8548**	0.8159	0.7627

影、且情感表现强度也有所减弱；而惊讶、恐惧的编辑结果中, 人物眼部出现显著的纹理失真。

结合图6的结果可以看出, 若不使用分类器, 则人物情感编辑的结果会产生瑕疵。首先, 对于第一列、第二列的开心与悲伤的表情编辑, 可以看出人物面部红润的特征在取消分类器后被削弱, 眼部出现黑色的晕影、并且情感表现强度也有所减弱；而对于惊讶、恐惧以及蔑视的编辑结果, 可以看出其中人物眼部出现了显著的纹理失真。这初步地验证了在多人的情感编辑任务使用统一的编辑向量, 将会在特定人物的情感编辑结果中产生瑕疵。同时, 有无分类器的定量实验结果如表2a所示, 可见引入分类器能带来更好的情感编辑效果。

取消分类器后的异常编辑结果, 可能是由于随着人物数目的增加, t-GAN 隐空间中的图像隐编码分布情况也逐渐复杂, 因此不同人物的不同情感的数据分布并不一致, 导致总体的情感编辑方向不能适用于每个人物。本文采取 t-SNE [45] 处理的方式, 分别选取 M003、W011 作为单个人物研究不同情感人脸的隐编码分布。其中, 每一种颜色代表一种情感。由此可见对单个人物而言, 同一种情感的数据呈现团簇状的分布；但是对于 M003、W011 而言, 不同情感的团簇分布模式有所差异, 因此编辑向量也会有所差异。为了更清晰展示这种差异, 本文选取了 M003、W011 中性情感以及高兴情感的隐编码进行 t-SNE 降维可视化, 如图7最右侧所示。其中紫色点、红色点分别代表 M003 中性情感和高兴情感人脸纹理的隐编码样本, 青色点、蓝色点分别代表 W011 中性情感和高兴情感人脸纹理的隐编码样本。由此可见, 不同人物由中性到高兴情感的编辑向量差异较大：尽管编辑方向基本一



(a) (b) (c)

图 7 (a)M003 不同情感人脸隐编码分布 (b)W011 的不同情感人脸隐编码分布 (c)M003、W011 的中性、高兴情感人脸隐编码分布

Figure 7 (a) The distribution of latent codes of faces with different emotion of M003 (b) The distribution of latent codes of faces with different emotion of M003 (c) The distribution of latent codes of faces with neutral and happy emotion.

图 8 W011 同时编辑几何和纹理（第一行）、只编辑几何（第二行）、只编辑纹理（第三行）的结果

Figure 8 The results of editing both shape and texture (Row 1), editing shape (Row 2), editing texture (Row 3) for subject W011.

致（M003 的、W011 的编辑方向分别由橙色、绿色箭头标出），但实现情感编辑隐空间编码需要移动的距离差异较大, M003 由中性到高兴情感的编辑距离明显大于 W011。因此使用统一的情感编辑方向进行隐空间编码的操作, 可能导致部分人物隐编码移动不合理, 进而导致编辑效果变差。

只编辑几何或者纹理 本部分选取 W011 作为研究对象, 进行每种情感的最大强度编辑, 如图8所示。其中列对应不同情感, 第一、二、三行分别为同时编辑纹理和几何、只编辑几何、只编辑纹理的结果。可见第二行中不同情感的编辑强度不明显, 而第三行中惊讶、害怕等情感编辑结果的眼部形状与目标情感强度不匹配。同时, 只编辑几何或者纹理的定量结果如表2b所示, 可见同时编辑人脸纹理、人脸几何才能得到更好的情感编辑效果。

不使用平滑化 本文模型中为保持情感编辑结果的帧间连续性, 使用了平滑化模块。这种连续性在视频中更容易体现出来, 因此该部分的消融实验结果主要参考本文补充材料中平滑化消融实验部分的视频。

5 总结与未来工作

本文在前人情感编辑相关工作的基础上做出了通用性层面的扩展与改进。本文通过设计纹理与几何解耦的方式, 并引入了基于人物分类器的情感向量编辑求解方法、基于 BFM 条件的人脸几何编辑网络等方法与模块, 使得情感编辑模型具有训练集已见人物、未见人物以及未见位姿三个层次的通用性。实验结果表明, 本文提出的模型从可视化以及定量指标的角度, 优于已有的可以应用到多人的情感编辑模型；消融实验也验证了分类器、解耦同时编辑等方法的有效性。针对目前研究中仍然存在的 t-GAN 隐空间多属性耦合的问题, 未来研究可以考虑寻求效果更好的情感编辑方式；同时未来的研究也可基于本文工作优化模型架构, 实现端到端的训练模式。

参考文献

1 Ye Z, Sun Z, Wen Y H, et al. Dynamic neural textures: Generating talking-face videos with continuously controllable expressions. arXiv preprint arXiv:2204.06180, 2022.

2 Ji X, Zhou H, Wang K, et al. Audio-driven emotional video portraits. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14080-14089.

3 Sun Z, Wen Y, Lv T, et al. Continuously controllable facial expression editing for talking videos. In Submission, 2022.

4 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition.In: In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

5 A. Mollahosseini, B. Hasani and M. H. Mahoor, “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” in IEEE Transactions on Affective Computing, (vol. 10, 1 Jan.- March 2019), pp. 18–31, doi: https://doi.org/10.1109/TAFFC.2017.2740923

6 Ding H, Sricharan K, Chellappa R. Exprgan: Facial expression editing with controllable expression intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence: volume 32. 2018.

7 T. Zhang, X. Wang, X. Xu and C. L. P. Chen, ”GCB-Net: Graph Convolutional Broad Network and Its Application in Emotion Recognition,” in IEEE Transactions on Affective Computing, vol. 13, no. 1, pp. 379-388, 1 Jan.-March 2022, doi: 10.1109/TAFFC.2019.2937768.

8 G. Zhang, M. Yu, Y. -J. Liu, G. Zhao, D. Zhang and W. Zheng, ”SparseDGCNN: Recognizing Emotion from Multichannel EEG Signals,” in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3051332.

9 X. Wang, T. Zhang and C. L. P. Chen, ”PAU-Net: Privileged Action Unit Network for Facial Expression Recognition,” in IEEE Transactions on Cognitive and Developmental Systems, 2022, doi: 10.1109/TCDS.2022.3203822.

10 Choi Y, Choi M, Kim M, et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8789-8797.

11 Prajwal K, Mukhopadhyay R, Namboodiri V P, et al. A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020: 484-492.

12 Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

13 Paysan P, Knothe R, Amberg B, et al. A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. IEEE, 2009: 296-301.

14 Wang K, Wu Q, Song L, et al. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: European Conference on Computer Vision. Springer, 2020: 700-717.

15 Deng Y, Yang J, Xu S, et al. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0. 58

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

17 Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2287-2296.

18 Blanz V, Vetter T. A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. 1999: 187-194.

19 Cao C, Weng Y, Zhou S, et al. Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on

Visualization and Computer Graphics, 2013, 20(3): 413-425.
20 Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph., 2017, 36(6): 194-1.
21 Luming Ma, Zhigang Deng: Real-Time Facial Expression Transformation for Monocular RGB Video. Computer Graphics Forum 38(1): 470-481 (2019)
22 Garrido P, Zollhöfer M, Casas D, et al. Reconstruction of personalized 3d face rigs from monocular video. ACM Transactions on Graphics (TOG), 2016, 35(3): 1-15.
23 Sanyal S, Bolkart T, Feng H, et al. Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7763-7772.
24 Feng Y, Feng H, Black M J, et al. Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-13.
25 Guo J, Zhu X, Yang Y, et al. Towards fast, accurate and stable 3d dense face alignment. In: European Conference on Computer Vision. Springer, 2020: 152-168.
26 Yi R, Ye Z, Zhang J, et al. Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137, 2020.
27 Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: Audio-driven facial reenactment. In: European conference on computer vision. Springer, 2020: 716-731.
28 Vougioukas K, Petridis S, Pantic M. Realistic speech-driven facial animation with gans. International Journal of Computer Vision, 2020, 128(5): 1398-1413.
29 Karras T, Aila T, Laine S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-12.
30 Wang T C, Liu M Y, Zhu J Y, et al. Video-to-video synthesis. arXiv preprint arXiv:1808.06601, 2018.
31 Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8110-8119.
32 Abdal R, Qin Y, Wonka P. Image2stylegan: How to embed images into the stylegan latent space?. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4432-4441.
33 Tov O, Alaluf Y, Nitzan Y, et al. Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-14.
34 Tewari A, Elgharib M, Bharaj G, et al. Stylerig: Rigging stylegan for 3d control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6142-6151.
35 Geng Z, Cao C, Tulyakov S. 3d guided fine-grained face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9821-9830.
36 Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4690-4699.
37 Zhu X, Lei Z, Liu X, et al. Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 146-155.
38 Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. 2015: 3730-3738.
39 Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in' Real-Life' Images: detection, alignment, and recognition. 2008.
40 Bulat A, Tzimiropoulos G. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision. 2017: 1021-1030.
41 Zhang Z, Song Y, Qi H. Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5810-5818.
42 Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 2017, 30.
43 Savchenko A V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In: 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY). IEEE, 2021: 119-124.
44 Narvekar N D, Karam L J. A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Transactions on Image Processing, 2011, 20(9): 2678-2683.
45 Van der Maaten L, Hinton G. Visualizing data using t-sne. Journal of machine learning research, 2008, 9(11).
46 Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans. Advances in neural information processing

systems, 2017, 30.

A continuous emotional editing model for talking head video based on decoupling texture and geometry

Tian LV[1], Yu-Hui WEN[2,3*], Zhiyao SUN[1] & Yong-Jin LIU[1*]

1. *Tsinghua University, Beijing* 100084*, China*;
2. *School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044*, China*;
3. *Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing* 100044*, China*

* Corresponding author. E-mail: yhwen1@bjtu.edu.cn, liuyongjin@tsinghua.edu.cn

Abstract The emotional editing of talking head video is one of the hot research topics in computer vision and computer graphics, which aims at converting a person's talking video with neutral emotion into another talking video with target emotion. Current methods could not simultaneously consider high-resolution emotional editing, maintenance of 3D property of human face and adaptability for different persons. To address this problem, We propose the BFM (Basel Face Model) conditioned shape editing network as our shape-emotion editing module, which guarantee the feasibility of geometric editing in multi-person conditions. And we propose the subject-classifier-based textural emotional editing module, which could preserve high-fidelity facial texture in multi-person tasks. Our proposed method breaks the limitations of previous emotional editing method, which could only be applied for a specific person, or could not generate high-resolution result in multi-person conditions. The experiment shows that our model can achieve better clarity, identity preservation, quality of editing than previous multi-person emotional editing methods, and can also get reasonable result on unseen person, and even on unseen head pose. Meanwhile, experiment shows that our model can continuously control the intensity of emotional editing.

Keywords Emotional editing, 3D reconstruction, Deep learning, Computer vision, Neural network

Tian LV is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2022. His research interests include 3D computer vision, machine learning and computer graphics.

Yu-Hui WEN is an associate professor with the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. She received her bachelor's degree from Harbin Institute of Technology (HIT), and the Ph.D. degree in computer science and technology from University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2020. She was a postdoc in Tsinghua University. Her research interests include machine learning, computer graphics and human motion analysis.

Zhiyao SUN is a Ph.D student with Department of Computer Science and Technology, Tsinghua University. He received his B.Eng. degree from Tsinghua University, China, in 2021. His research interests include computer vision and computer graphics.

Yong-Jin LIU is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include affective computing, computer graphics and computer vision. He is a senior member of IEEE and ACM. For more information, visit https://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.html