

基于三维的人脸合成与替换系统实现

钟卓岑, 史泽奇

浙江理工大学信息学院, 浙江 杭州

收稿日期: 2022年4月16日; 录用日期: 2022年6月8日; 发布日期: 2022年6月15日

摘要

在保持人物身份的前提下对人脸合成, 也称为人脸替换, 目的将源图像中的人脸无缝且逼真地融合到目标图像。本文提出了一种自动化的人脸合成与替换系统, 利用带有标签的图像合成用户感兴趣的全新形象, 实现如虚拟试发、虚拟试衣的应用。系统的输入为单张人脸图像和查询文本, 其中查询文本用于搜索相关图像, 构建待替换的目标图像集合。为提高人脸合成效果, 本文首先通过人脸图像重建三维人脸模型, 在三维空间完成了人脸的姿态与表情迁移, 然后通过人脸图像分解估计环境光照, 实现了人脸图像的重光照, 最后通过图像融合技术实现人脸的图像融合。实验证明我们的方法可以合成较为真实的人脸图像。

关键词

人脸替换, 人脸合成

Realization of 3D-Based Face Synthesis and Replacement System

Zhuocen Zhong, Zeqi Shi

School of Information, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Apr. 16th, 2022; accepted: Jun. 8th, 2022; published: Jun. 15th, 2022

Abstract

Face synthesis, also known as face replacement, while maintaining the identity of the person, aims to seamlessly and realistically fuse the face in the source image to the target image. This paper proposes an automated face synthesis and replacement system, which uses labeled images to synthesize new images that users are interested in, and realizes applications such as virtual hair test and virtual fitting. The input of the system is a single face image and query text, where the query text is used to search for relevant images and construct a set of target images to be replaced. In

order to improve the effect of face synthesis, this paper firstly reconstructs a 3D face model through face images, completes the face pose and expression migration in 3D space, and then estimates the ambient lighting through face image decomposition to realize the re-illumination of face images, and finally realize the image fusion of the face through the image fusion technology. Experiments show that our method can synthesize more realistic face images.

Keywords

Face Replacement, Face Synthesis

Copyright © 2022 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着计算机技术和互联网技术的迅猛发展,出现了许多兼具实用性和趣味性的应用,改变了人们的生活方式。人脸作为人的外在特征,是计算机视觉与计算机图形学的重点研究对象。相较于人脸检测与识别,人脸的形象合成相关的技术更加开放。通过改变人物发型或职业,人物的形象会发生巨大的变化,带了许多实际的应用。例如,通过将人脸与发型合成,可以让人随时尝试各类造型,避免多次奔波理发店;通过将人脸融合到穿着不同服装的模特图像,可以实现虚拟试衣;在新闻报道或者全景地图拍摄时,接触人脸合成技术可以将人脸替换为通用人脸,可以起到保护隐私的作用。

近年来,人脸合成和替换的应用出现在各类娱乐软件,在一些娱乐软件中已经实现了将用户输入的人脸图片合成为其他有趣的新形象。但这些人脸合成通常应用在限在特定的形象风格中,存在泛化能力不足,无法根据用户的需求定制人脸形象的缺点。

随着计算机视觉和图形学技术的发展,人脸处理相关的技术也得到了突飞猛进的进展。由于技术的使用门槛相对较高以及技术本身的限制,普通的大众人群难以实现根据自己的想法合成特定的形象。为此,本文基于人脸检测、三维人脸重建、人脸重光照等与人脸图像处理相关的深度学习技术,结合 Python、PyQt,提出和实现了可根据用户需要合成相关人脸图像的综合解决方案。在本系统中,根据用户输入的文本内容,例如人物发型、职业、服装等关键字,通过互联网搜索相关人物图像并进行分析和处理,在保持输入人脸图像的身份的前提下完成人脸的无缝合成与替换。

2. 系统流程介绍

我们的系统将单张人脸图像和文本关键字作为输入,其中输入的人脸图像称为源图像。文本关键词通过在互联网中查找相关图像集合,并与源图像经过人脸处理模块分别处理,在保持输入人脸图像的身份的前提下完成人脸的无缝合成与替换,向用户展现全新的人物形象。

系统的整体框架流程图 1 所示。图中蓝色虚线框内为系统输入,最右侧为输出的部分图像。人脸图像和文本关键字作为系统输入,文本关键字用来查询并生成待替换的人脸图像集合,本文将输入的人脸图像称为源图像,待替换图像集合称为目标图像集合。在系统中,要处理的内容主要包括:图像搜索、人脸检测、人脸姿态估计、人脸属性分割、三维人脸重建、光照估计与迁移、纹理映射和人脸图像融合等。系统的输出为替换人脸后的图像集合,保持了源图像人脸的身份和目标图像集合的背景。

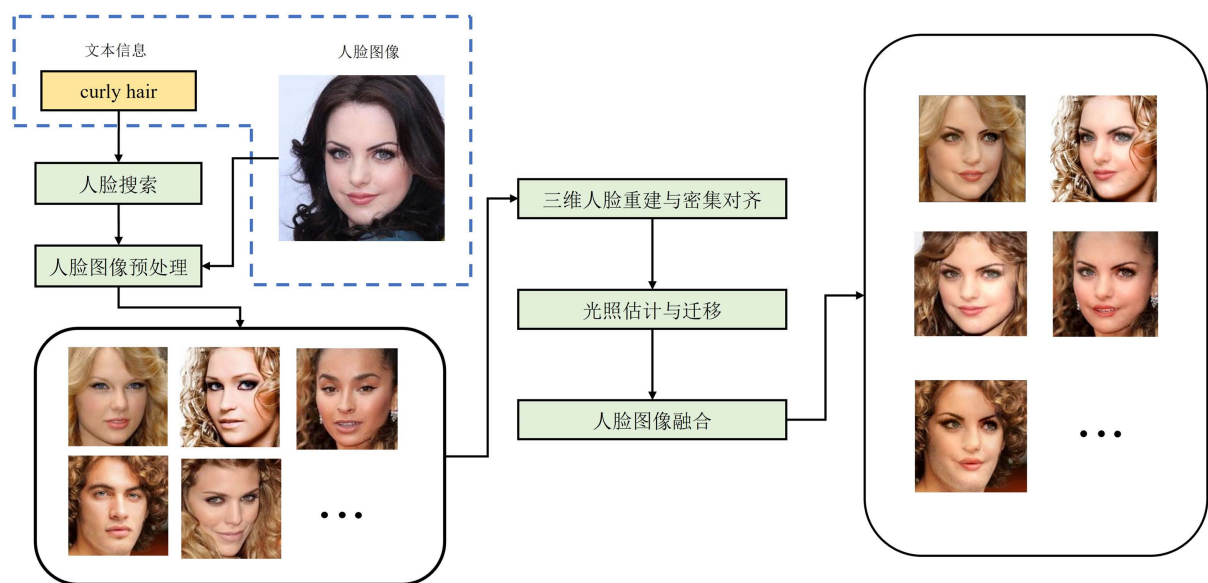


Figure 1. Automatic face replacement pipeline
图 1. 自动人脸替换流程图

3. 系统模块介绍

如图 2 所示, 根据不同的任务, 本文将系统的核心模块分为输入模块、图像搜索模块、图像预处理模块、三维人脸重建模块、光照估计与迁移模块和人脸合成与替换模块。

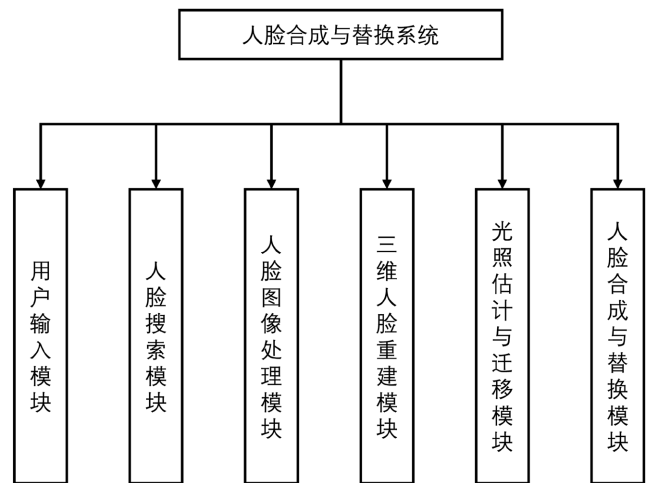


Figure 2. System module
图 2. 系统模块

- 1) 用户输入模块: 用户输入模块用来向用户提供输入接口, 接收用户输入, 包括文本输入和图片路径选择。
- 2) 人脸搜索模块: 该模块负责从本地库或互联网中查找与输入文本关键字相关的人脸图像。系统运行后, 人脸搜索模块从前端用户界面获取文本进行判断。若文本不为空, 则从本地库或缓存图片库中查找与该文本具有相符标签的图像集合。若不存在这类的图像, 则通过搜索引擎接口从互联网中查询相关图像, 缓存到本地库后返回该图像集合。具体流程如图 3 所示。

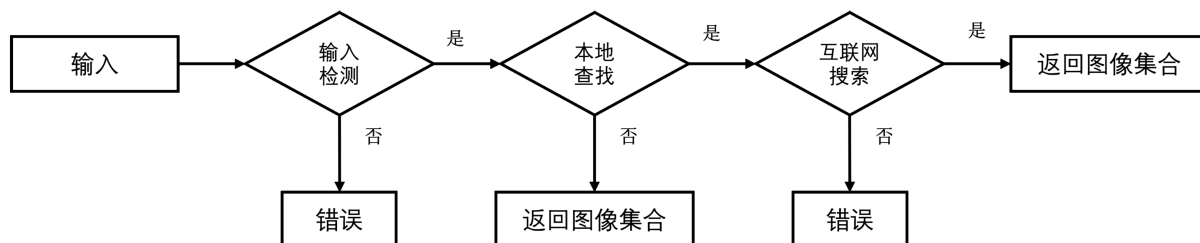


Figure 3. Face search process

图 3. 人脸搜索流程

3) 图片预处理模块：图片预处理模块用来完成对源图像和目标图像集中人脸图像的预处理，包括人脸检测、人脸姿态估计和替换区域的 mask 生成。如图 4 所示，首先通过人脸检测算法查找图像的人脸区域，人脸检测算法用来过滤非人脸图像和人脸分辨率相差较大的图像，并裁剪人脸区域，同时输出的 5 个人脸关键点。5 个人脸关键点用来粗略估计人脸姿态，通过人脸姿态估计算法，获得人脸姿态的欧拉角表示，用来加速去除与源图像人脸姿态相差较大的目标图像集中的图像。裁剪的人脸区域经过人脸分割算法，完成人脸属性的分割，并将生成的人脸 mask 映射到原始图像，获得在原始图像分辨率下的人脸 mask。

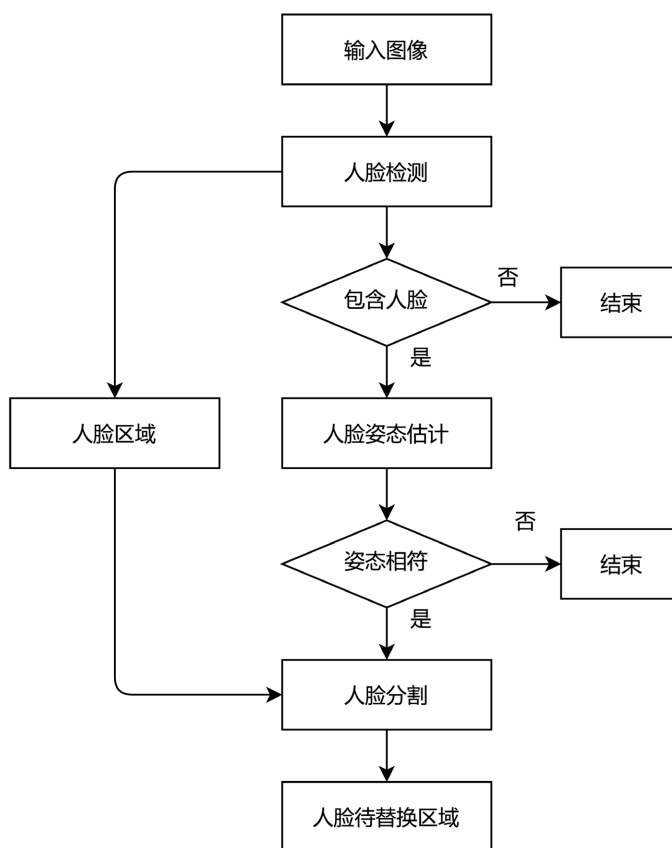


Figure 4. Image preprocessing process

图 4. 图片预处理流程

4) 人脸重建模块：人脸重建模块用于生成与人脸图像相对应的三维人脸模型。如图 5 所示，人脸重建模块包含了三个子模块：三维人脸参数回归模块、三维人脸模型生成模块和渲染模块。在三维人脸参

数回归模块中, 分别获得源图像和待替换的目标人脸图像集合对应的三维形变模型(3D Morphable Model, 3DMM) [1]系数, 通过将源图像的三维形变模型表情系数和姿态系数修改为目标图像集合人脸的表情系数和姿态系数, 并保持形状系数不变, 可实现三维人脸的表情与姿态迁移。三维人脸模型生成模块用来根据形变模型系数生成三维人脸模型。渲染模块根据姿态系数定义相机位置, 将三维人脸模型渲染为二维图像。

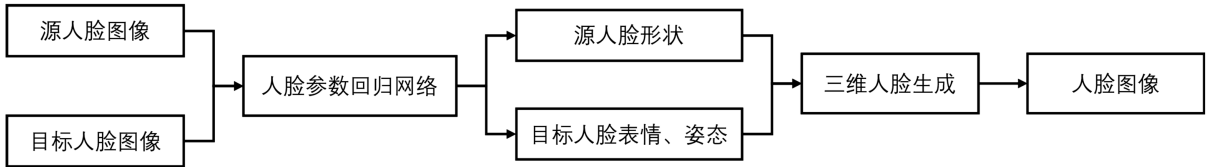


Figure 5. Face reconstruction process
图 5. 人脸重建流程

5) 光照处理模块: 基于三维形变模型方法可以重建人脸的形状和表情, 但对于人脸的纹理重建效果不好, 只能生成较为光滑的人脸纹理, 无法恢复胡须、皱纹等高频细节。因此, 文本将原始图像中的人脸颜色作为纹理映射到三维人脸模型, 但原始图像的人脸颜色包含了光照等信息, 并不是人脸真正的颜色。不同的图像所处的环境光照不同, 也会呈现出不同的人脸光影效果。光照处理模块用于完成人脸图像的光照估计与迁移任务, 生成在目标图像光照下的人脸图像, 流程如图 6 所示。

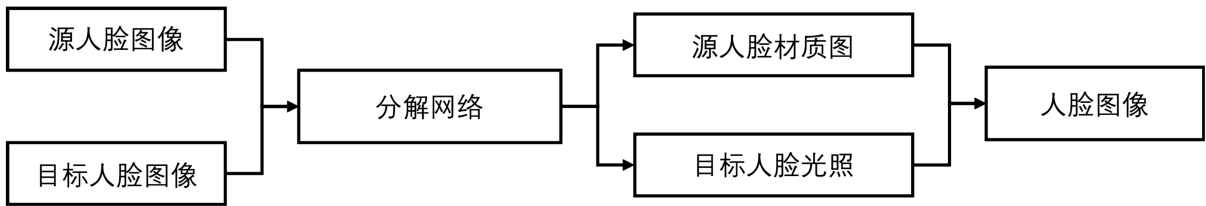


Figure 6. Relighting process
图 6. 重光照流程

6) 人脸替换模块: 人脸替换模块用来实现最终的人脸融合与替换, 将重光照后的源图像映射到三维模型中, 并将渲染后的人脸融合到目标图像的待替换区域。

4. 系统模块实现

4.1. 人脸图像处理

人脸图像预处理模块包括了人脸检测、人脸姿态估计和人脸分割, 用来过滤非人脸和与输入人脸图像姿态相差过大的图像。本文使用多任务级联卷积神经网络[2] (MTCNN)算法完成人脸检测和姿态估计。MTCNN 算法使用多任务联合训练的方式, 采用了候选框加分类器的思想, 将人脸检测与人脸关键点检测结合在了一起。如图 7 所示, MTCNN 算法首先通过对输入的图像进行缩放操作来建立人脸图像金字塔, 用来检测不同尺度的人脸图像, 并依次通过提议网络 P-Net、优化网络 R-Net 和输出网络 O-Net, 由粗到细地预测人脸外接框和 5 个人脸关键点的位置, 最后根据人脸外接框进行裁剪获得人脸区域。

根据 5 个人脸关键点和三维人脸通用模型的对应关系, 可粗略估计出图像中的人脸姿态, 用来去除与输入人脸姿态相差过大的图像。物体姿态是指物体相对于相机的方向, 本文使用欧拉角(俯仰角 pitch、偏航角 yaw 和翻滚角 roll)来表示人脸姿态。二维图像与三维通用模型对应位置如表 1 所示。

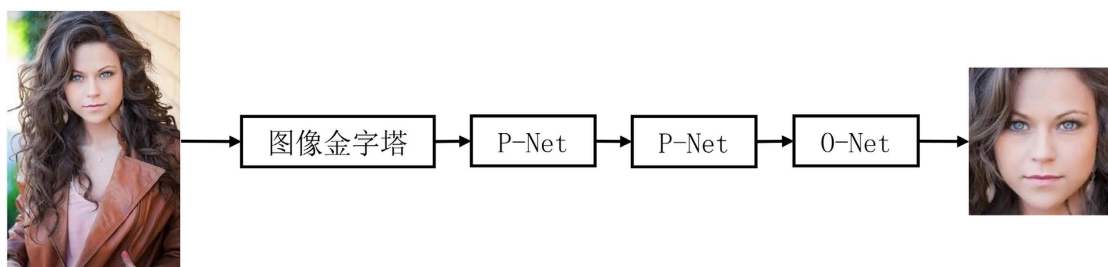


Figure 7. MTCNN algorithm process

图 7. MTCNN 算法流程

Table 1. Key point corresponding position

表 1. 关键点对应位置

关键点名称	世界坐标位置
鼻子	(0.0, 0.0, 0.0)
左眼睛	(-165.0, 170.0, -115.0)
右眼睛	(165.0, 170.0, -115.0)
左嘴角	(-150.0, -150.0, -125.0)
右嘴角	(150.0, -150.0, -125.0)

三维人脸到二维人脸的映射涉及到空间坐标系的转化, 由三维人脸所在的世界坐标系到二维图像的像素坐标系的映射关系如式(1)所示:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & u_0 \\ 0 & \frac{1}{dy} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{f}{z_c} & 0 & 0 & 0 \\ 0 & \frac{f}{z_c} & 0 & 0 \\ 0 & 0 & \frac{1}{z_c} & 0 \end{bmatrix} \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \quad (1)$$

其中 u 和 v 为人脸图像点 p 所在的像素坐标系的坐标, u_0 和 v_0 为图像平面中心, dx 和 dy 为点 p 在 u 轴和 v 轴上的单位距离, f 为相机焦距, z_c 为相机坐标系下的 Z 轴分量, R 为旋转矩阵, t 为平移向量, x_w , y_w 和 z_w 为点 p 对应的世界坐标系下的三维人脸模型。通过 OpenCV 图像库提供的 solvePnP() 函数, 可获得旋转矩阵 R 与平移向量 t , 并将旋转矩阵 R 转为欧拉角表示[3]。图 8 展示了人脸姿态估计的结果, 测试图片来自 CelebAMask-HQ [4] 人脸数据集。



Figure 8. Pose estimation result

图 8. 姿态估计结果

对于无约束的二维人脸图像, 可能由于姿态、头发、装饰等造成人脸区域存在遮挡, 对最终的人脸合成效果产生较大的影响。因此, 在图像预处理阶段, 需要对人脸区域与其他区域做出区分, 仅对人脸图像的面部区域进行处理, 以更好的完成人脸合成与替换的任务。本文通过修改 BiSeNet [5] [6] 网络在数据集 CelebAMask-HQ [4] 中训练, 将人脸图像分割为面部皮肤、鼻子、眼睛、眉毛、嘴巴、头发、眼镜等区域。人脸分割可以防止由于人脸遮挡问题带来的错误的结果, 同时根据用户需要合成佩戴眼镜等其他装饰物, 图 9 展示了在人脸分割和生成的人脸 mask 的结果。



Figure 9. Face segmentation result
图 9. 人脸分割结果

4.2. 人脸重建与密集对齐

二维的人脸图像本身所能表达的人脸信息含量有限, 在三维空间对人脸进行处理, 可以使合成的人物形象更加真实和丰富。三维形变模型构建的人脸具有相同的拓扑结构, 模型中的每一个关键点坐标都具有相同的语义信息。本文通过人脸图像重建完成人脸图像在三维空间中的密集人脸对齐。

传统的三维形变模型是在人脸数据的基础上统计人脸的形状和纹理, 任意的一个人脸模型看作是形状模型和纹理模型的线性组合, 可以由 m 个人脸模型通过线性组合得到。与之类似, 本文采用的模型利用形状基向量和表情基向量的线性组合来描述三维人脸空间, 用来拟合生成在不同状态下的具有不同表情的无约束三维人脸。一个全新的人脸模型表示为如下形式:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp} \quad (2)$$

其中 S 为恢复的三维人脸模型, $\bar{S}$ 为平均人脸的形状, A_{id} 和 A_{exp} 为人脸和表情的主成分, α_{id} 和 α_{exp} 为形状参数和表情参数, 通过拟合人脸图像的形状和表情参数, 即可获得该人脸图像对应的三维人脸模型。

对于人脸的不同姿态, 将三维人脸 S 通过弱视透视投影到二维的平面, 可得:

$$V_{2d}(p) = f * Pr * R * (\bar{S} + A_{shp}\alpha_{shp} + A_{exp}\alpha_{exp}) + t_{2d} \quad (3)$$

$V_{2d}(p)$ 表示人脸重建和投影函数, 将三维人脸顶点投影到二维平面, 获得其在二维平面的位置坐标, f 为缩放因子, Pr 为 2×3 的正交投影矩阵 $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, R 和 t_{2d} 分为旋转矩阵和平移向量。

本文参考 3DDFA [7] 算法, 设计实现了一个卷积神经网络回归人脸的形状系数、表情系数和姿态系数, 并通过改变源图像的表情系数和姿态系数, 渲染了在目标图像表情和姿态下的人脸图像, 用于后续

的人脸替换。图 10 展示了表情和姿态迁移的结果, 其中第一行为输入的源人脸图像与目标人脸图像, 第二行为重建结果。



Figure 10. Face reconstruction result
图 10. 人脸重建结果

4.3. 光照估计与迁移

不同的人脸图像所处的环境光照不同, 人脸图像光照迁移对人脸更加逼真地融入到目标光照环境有着重要的作用。本文在假设人脸为 Lambert 模型的前提下, 使用球面谐波模型来近似表示人脸图像的光照环境。大量实验研究表明, Lambert 表面在不同光照条件下的图像处于一个低维的子空间内, 因此许多研究尝试使用统计分析或解析法对不同光照变化下的图像进行建模, Basri [8]等人提出的球谐光照模型不再需要估计光源的方向, 大大简化了对光照的估计。根据球面谐波理论, 将辐照度 E 在频域中用球面谐波函数展开:

$$E(n) = \sum_{l=0}^{\infty} \sum_{m=-l}^l A_l L_l^m Y_l^m(n) \approx \sum_{l=0}^2 \sum_{m=-l}^l A_l L_l^m Y_l^m(n) \quad (4)$$

其中 A_l ($A_0 = \pi, A_1 = 2\pi/3, A_2 = \pi/4$) 是 Lambertian 反射的球面谐波系数, Y_l^m 是球面谐波函数, L_l^m 是入射光线的系数。

由朗伯定律可以知道, 对于图像上的每一个像素点, 都具有球面上相同法向量处的光照强度。在假设人脸为朗伯凸表面的前提下, 给定一副人脸纹理的图像 I , 对于图像上的每一个像素 (x, y) , 下面等式几乎都成立:

$$I(x, y) = \rho(x, y) E(n) \quad (5)$$

其中 n 为人脸图像中点 (x, y) 的法线方向, $\rho(x, y)$ 为人脸的反照率, 为常数。将辐照度的球面谐波方程代入上式子, 得:

$$I(x, y) \approx \sum_{l=0}^2 \sum_{m=-l}^l \rho(x, y) A_l L_l^m Y_l^m(\alpha(x, y), \beta(x, y)) \quad (6)$$

因此, 对于 Lambert 模型下的人脸图像, 可将其分解为与光照无关的反照率图, 法向图和用球谐系数表示的环境光照。基于这个分析, 本文参考 SfSNet [9]设计实现了一个编码器-解码器的沙漏型卷积神经网络, 用于从单张图像分解人脸几何、材质与光照环境, 并重新渲染在目标光照下的人脸图像。图 11 展示了重光照的结果。



Figure 11. Relighting result
图 11. 重光照结果

5. 系统界面实现

由于 Python 内置的 Tkinter 库功能较少, 无法高效地完成专业的 GUI 开发, 本文使用了 PyQt5 替代 Tkinter 完成界面开发。PyQt 是 Qt 的 Python 绑定版, 由 Riverbank Computing Limited 开发, 将 Python 编程语言和 Qt 库相融合, 用于创建 GUI 应用程序的工具包[10], 它是 QT 跨平台框架中受欢迎的 Python 绑定之一。PyQt 实现了一系列 Python 模块, 由 620 多个类和构成。利用了信号和槽的机制完成不同对象之间的通信, 可方便地创建可复用的模块。同时, 它还是一个多平台工具包, 可以在几乎所有主要操作系统包括 Linux、Windows 和 Mac 上运行。

系统的主界面如图 12 所示, 主要分为了两个区域: 文本图片输入区和结果展示区。输入区域可输入查找的关键字和读取源人脸图像并展示, 输出区域包含了个板块, 分别为目标图像展示板块、人脸合成结果展示板块。



Figure 12. System interface
图 12. 系统界面

6. 结果展示

图 13 为本文系统在输入单张人脸图像和查询文本的部分结果, 查询文本分别为“blond hair”, “lawyer”和 “student”, 第一列为输入人脸图像, 后四列为输出的合成人脸图像。可以看出本文的方法可以实现较好的人脸合成效果。



Figure 13. Face synthesis and replacement result
图 13. 人脸合成与替换结果

参考文献

- [1] Blanz, V. and Vetter, T. (1999) A Morphable Model for the Synthesis of 3d Faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, 8-13 August 1999, 187-194. <https://doi.org/10.1145/311535.311556>
- [2] Zhang, K., Zhang, Z., Li, Z., et al. (2016) Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, **23**, 1499-1503. <https://doi.org/10.1109/LSP.2016.2603342>
- [3] Kaehler, A. and Bradski, G. (2016) *Learning OpenCV 3: Computer Vision in C++ with the OpenCV library*. O'Reilly Media, Inc., Newton.
- [4] Lee, C.H., Liu, Z., Wu, L., et al. (2020) Maskgan: Towards Diverse and Interactive Facial Image Manipulation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 5549-5558. <https://doi.org/10.1109/CVPR42600.2020.00559>
- [5] Yu, C., Wang, J., Peng, C., et al. (2018) Bisenet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., (Eds.), *Computer Vision-ECCV 2018. ECCV 2018. Lecture Notes in Computer Science*, Heidelberg, Cham, 325-341. https://doi.org/10.1007/978-3-030-01261-8_20
- [6] Fan, M., Lai, S., Huang, J., et al. (2021) Rethinking Bisenet for Real-Time Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 9716-9725. <https://doi.org/10.1109/CVPR46437.2021.00959>
- [7] Zhu, X., Liu, X., Lei, Z., et al. (2017) Face Alignment in Full Pose Range: A 3d Total Solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 78-92. <https://doi.org/10.1109/CVPR46437.2021.00959>
- [8] Frolova, D., Simakov, D. and Basri, R. (2004) Accuracy of Spherical Harmonic Approximations for Images of Lambertian Objects under Far and Near Lighting. In: Pajdla, T. and Matas, J., (Eds.), *Computer Vision-ECCV 2004. ECCV 2004. Lecture Notes in Computer Science*, Springer, Heidelberg, 574-587. https://doi.org/10.1007/978-3-540-24670-1_44
- [9] Sengupta, S., Kanazawa, A., Castillo, C.D., et al. (2018) Sfsnet: Learning Shape, Reflectance and Illuminance of Faces in the Wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6296-6305. <https://doi.org/10.1109/CVPR.2018.00659>
- [10] Willman, J.M. (2022) *Creating Custom Widgets. Beginning PyQt*. Apress, Berkeley, 341-358. https://doi.org/10.1007/978-1-4842-7999-1_12