文章编号：2095-2163(2023)07-0071-05 中图分类号：TP391.41 文献标志码：A

基于 SMPL 的人物视频生成算法

范沈伟，李国平，王国中
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：人物视频的生成需要人物外观和人物运动两部分信息。现有算法大都使用人物关键点在 2D 平面内的运动趋势作为运动信息，生成视频的视角只能是固定的。为了生成多视角的人物视频，提出了基于 SMPL 的人物视频生成算法。算法首先使用 NERF 配合 SMPL 人体模型对人物进行 3D 建模，获取人物外观信息。随后，提取驱动视频中的 SMPL 人体模型参数作为人物的运动信息，驱动静态的人物 3D 模型。最后，使用 NERF 中的体渲染技术将人物 3D 模型映射到 2D 平面内，得到最终的人物视频。得益于 SMPL 模型的特点，生成的视频不仅可以任意切换视角，还可以任意修改人物的体型。

关键词：SMPL；NERF；视频生成

Human video generation algorithm based on SMPL

FAN Shenwei, LI Guoping, WANG Guozhong
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 The generation of the human video requires two parts of information: human appearance and human movement. Most existing algorithms treat human key points in 2D plane as the movement information, so the view of the generated video can only be fixed. This paper proposes a human video generation algorithm based on SMPL to generate multi-view human video. Firstly, the algorithm uses NERF cooperated with SMPL to get human appearance information by building a 3D human model. Then, the algorithm extracts SMPL parameters from the driving video as human movement information. Finally, in order to get the generated human videos, the algorithm uses volume rendering technique in NERF to map the 3D human model to 2D plane. Due to the property of SPML, the generated video can change not only the view, but also the shape of human.

【Key words】 SMPL; NERF; video generation

0 引 言

在虚拟现实以及动画的制作中，人物视频的生成是一个重要的研究课题。由于人物的外观包含大量的细节并且人物的运动由身体各个部分的运动状态决定，所以想要生成高质量的人物视频是一项具有挑战性的工作。

Vid2Vid[1] 提出了一种基于条件对抗生成网络 CGAN[2] 的视频生成模型。模型能够使用 2D 关键点驱动关键帧生成视频。一方面，模型预测当前帧相对于前一帧的光流，通过扭曲前一帧得到当前帧；另一方面，模型使用 CGAN 来生成无法使用光流扭曲得到的部分。

一个 Vid2Vid 模型只能对应于一个人物，Few-shot Vid2Vid[3] 在 Vid2Vid 的基础上实现了一个模型对应多个人物。模型引入了生成器参数生成模块。其中的生成器参数不再是固定的，而是根据输入的人物图片来生成对应的生成器参数。相比于 Vid2Vid，Few-shot Vid2Vid 训练更加困难，需要庞大的数据集，并且训练成本很高。

Vid2Vid[1] 和 Few-shot Vid2Vid [3] 在训练以及生成视频的过程中需要使用预训练的关键点提取网络获取关键点标签信息，Monkeynet[4] 提出了一种端到端的无监督视频生成模型，不再需要事先提取关键点。模型使用了自编码器结构，首先使用帧与帧之间关键点的位移来预测帧与帧之间的光流，然后根据光流扭曲自编码器的隐式编码来得到生成的帧。

基金项目：国家重点研发计划（2019YFB1802700）。

作者简介：范沈伟（1995-），男，硕士研究生，主要研究方向：计算机视觉；李国平（1974-），男，博士，高级工程师，硕士生导师，主要研究方向：音视频编码、智能媒体处理、机器学习；王国中（1962-），男，博士，教授，博士生导师，主要研究方向：音视频编码、图像处理。

通讯作者：李国平 Email：liguoping@ sues.edu.cn

收稿日期：2022-08-15

由于简单的关键点之间的位移并不能很好地表示关键点附近像素点的运动趋势。所以除了预测关键点之外,FOMM[5] 还增加了雅可比矩阵的预测。将关键点及关键点附近像素点的运动趋势看作是 2D 平面内的仿射变换。由于获取了更加准确的光流,FOMM[5] 相比于 Monkeynet[4] 生成的视频质量更好。

国内文献[6-9]在国外工作的基础上提出了几种不同的生成网络结构进行人物视频的生成。

上述算法都使用人物 2D 图像作为外观信息,人物关键点在 2D 平面内的运动趋势作为运动信息来进行视频的生成,只能生成固定视角的视频。为了生成多视角的人物视频,本文引入了 3D 信息取代 2D 信息。区别于使用贴图的方式进行人物 3D 建模[10-12],本文提出了一种结合 SMPL[13] 以及 NERF[14] 的方法获取人物 3D 模型。考虑到 3D 模型由 SMPL 获得,因此利用 VIBE[15] 得到的 SMPL 参数可以对其进行驱动,以此完成多视角人物视频的生成。

1 相关技术

1.1 NERF

神经辐射场(NERF)是一种通过稀疏的不同视角的图片,对静态三维物体进行重建的技术。过程中使用一个神经网络(多层感知机 MLP)来表示一个静态的三维物体。研究输入是观测视角 (θ, ϕ) 和三维坐标点 (x,y,z),输出是体密度 σ 以及三维坐标点的颜色 c。使用输出 (σ,c) 进行体渲染,就能得到三维物体在不同视角下的二维图像。

体渲染首先根据观测视角和相机参数确定观测点以及图像平面。以观测点为起点,以图像平面每个像素的中心为目标发射射线。在近平面和远平面之间,将射线均分成 n 段。每段的长度记为 δ。在每一个小段中进行随机采样,并送入 NERF 得到对应三维坐标点的 σ,c。根据式(1)进行积分得到射线与图像平面交点处像素的值:

$$
\begin{aligned}
&C(t) = \int T(t)\sigma(t)c(t)\mathrm{d}t \approx \sum_{i=1}^{n} T_i c_i \alpha_i \\
&r(t) = o + td \\
&T_i = \prod_{j=1}^{i-1}(1-\alpha_j) \\
&\alpha_i = 1 - \exp(-\sigma_i \delta_i)
\end{aligned} \tag{1}
$$

1.2 SMPL

SMPL 模型是一个裸体的 3D 人体模型,具有 N =6 890 个顶点和 K = 23 个关节。由 $\vec{\beta}$ 和 $\vec{\theta}$ 两个参数来分别控制模型的体型和姿势。$\vec{\beta}$ 的维度 $|\vec{\beta}|$ 为 10,$\vec{\theta}$ 的维度 $|\vec{\theta}| = 3 * K + 3 = 72$,每个关节对应一个三维旋转向量,外加一个全局旋转向量。该过程可由如下公式进行描述:

$$
\begin{aligned}
&T_P(\vec{\beta},\vec{\theta}) = \bar{T} + B_s(\vec{\beta}) + B_p(\vec{\theta}) \\
&B_s(\vec{\beta}) = \boldsymbol{S}\vec{\beta} \\
&B_p(\vec{\theta}) = \boldsymbol{P}(R(\vec{\theta}) - R(\vec{\theta}^{\,*}))
\end{aligned} \tag{2}
$$

其中,$\bar{T} \in \mathbb{R}^{3N}$ 是 SMPL 模型的平均模板,保持 T-pose。当 $\vec{\beta}$ 和 $\vec{\theta}$ 变化时,模型顶点会在 $\bar{T}$ 的基础上发生偏移。偏移量的大小由 B_s 和 B_p 函数计算得到;$\boldsymbol{S}$ 表示一个 $3N * |\vec{\beta}|$ 的矩阵;$\boldsymbol{P}$ 表示一个 $3N * 9K$ 的矩阵;R 表示罗德里格斯公式,将三维旋转向量变换为 $3 * 3$ 的旋转矩阵。关节点的三维坐标由式(3)计算得到:

$$
J(\vec{\beta}) = \Im(\bar{T} + B_s(\vec{\beta})) \tag{3}
$$

其中,$\Im$ 表示一个 $(K+1) * N$ 的矩阵。

进一步地,研究推得的数学公式可写为:

$$
\begin{aligned}
&M(\vec{\beta},\vec{\theta}) = \boldsymbol{W}(T_P(\vec{\beta},\vec{\theta}),J(\vec{\beta}),\vec{\theta},\boldsymbol{W}) \\
&\bar{t}_i^{\,'} = \sum_{k=0}^{K} w_{k,i} G_k'(\vec{\theta},J)\bar{t}_i \\
&G_k'(\vec{\theta},J) = G_k(\vec{\theta},J)G_k(\vec{\theta}^{\,*},J)^{-1} \\
&G_k(\vec{\theta},J) = \prod_{j\in A(k)} \begin{pmatrix} R_j & J_j \\ 0 & 1 \end{pmatrix}
\end{aligned} \tag{4}
$$

得到了 $T_P(\vec{\beta},\vec{\theta}),J(\vec{\beta}),\vec{\theta}$ 之后,SMPL 模型使用混合矩阵 $\boldsymbol{W}$ 作为权重,通过混合函数 $\boldsymbol{W}$ 得到最终的顶点。式(4)中,混合矩阵 $\boldsymbol{W}$ 是 $N * (K+1)$ 的矩阵,表示每个关节对每个顶点的影响。$A(k)$ 表示前 k 个关节的集合。$G_k(\vec{\theta},J)$ 表示关节 k 相对于世界坐标系的变换,$G_k'(\vec{\theta},J)$ 表示关节 k 相对于 T-pose 的变换。

2 本文方法

2.1 算法流程

本文提出的基于 SMPL 的人物视频生成算法主要包括 3 个模块,分别是:视频前后景分割模块、NERF 三维重建模块、以及 SMPL 参数预测模块。算法整体流程如图 1 所示。

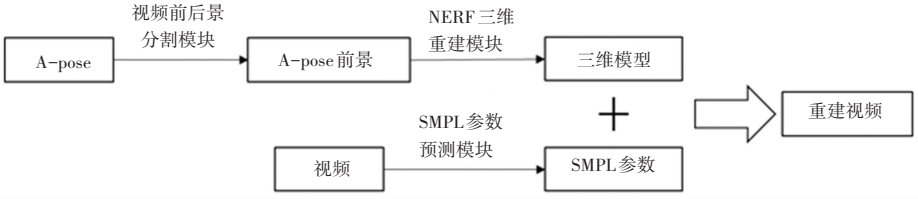


图 1　算法流程图

Fig. 1　Overview of the method

算法首先需要对视频中的人物完成三维建模。使用单目摄像头固定好拍摄位置，拍摄一段目标人物保持 A-pose，原地 360°旋转的视频，如图 2 所示。为了不让背景影响三维重建的过程，使用视频前后景分割模块[16]将目标人物作为前景提取出来，然后送入 NERF 训练得到三维模型。

有了目标人物的三维模型，要生成人物视频，还需要获得一组 SMPL 模型参数作为运动信息。SMPL 参数由 SMPL 参数预测模块 VIBE[15]从一段视频中提取。

图 2　A-pose 示意图

Fig. 2　A-pose explanation

2.2　人物三维重建

2.2.1　三维坐标转换

NERF 三维重建模块需要将 A-pose 图像序列转化为相应的三维模型，详见图 2。由于在 A-pose 获取的过程中，需要人物在原地 360°转圈。在没有特殊设备的情况下，人的姿态不可避免地会有改变。而 NERF 适用于静态场景的重建，人物姿态的变化会导致重建质量的下降。

为此，本文将图像中人物不同姿态所对应的 3D 空间坐标使用 SMPL 模型统一变换为 SMPL 平均模板中的姿态 T-pose，如图 3 所示。接着，再使用 NERF 进行重建。在式(4)基础上，继而可推得：

$$v_i' = \sum_{k=0}^{K} w_{k,i} G_k'(\vec{\theta}, J)\left(v_i + b_{s,i}(\vec{\beta}) + b_{p,i}(\vec{\theta})\right) \tag{5}$$

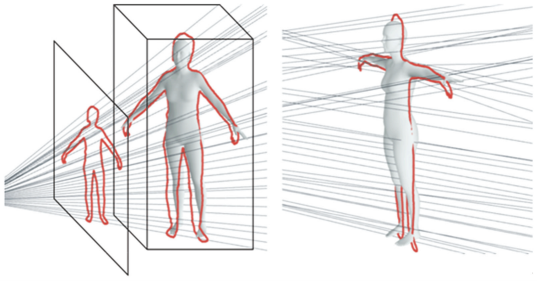


图 3　A-pose 到 T-pose 的空间坐标转换示意图

Fig. 3　Coordinate transformation from A-pose to T-pose

空间坐标变换方法参考文献[10]，见式(6)：

$$v_i + b_{s,i}(\vec{\beta}) = \left(\sum_{k=0}^{K} w_{k,i} G_k'(\vec{\theta}, J)\right)^{-1} v_i' - b_{p,i}(\vec{\theta}) \tag{6}$$

其中，v 表示 SMPL 模型顶点，$b_{s,i}$ 和 $b_{p,i}$ 分别是顶点 i 的体型函数以及姿势函数。式(6)的左边是 T-pose 的人物顶点坐标，右边的 v_i' 是 A-pose 的人物顶点坐标。

2.2.2　三维重建模块整体结构与损失函数

去除背景的 A-pose 作为输入，经过 SMPL 参数预测模块后得到弱透视相机参数以及 SMPL 参数。根据相机参数进行采样。以相机为起点发射射线，射线的方向穿过每个像素的中心。在视频前后景分割时可以得到人物的轮廓掩膜，只有在轮廓内部的射线才得到保留。另外，根据 SMPL 模型的顶点信息可以得到一个包围 3D 人体的立方体。三维重建模块整体结构如图 4 所示。假如射线和立方体有两个交点，则在该条射线与立方体的 2 个交点之间在三维空间中进行采样。NERF 的采样分 2 次，第一次粗略采样，第二次在粗略采样的基础上再进行一次精细采样。2 次采样经过体渲染(volume rendering)后会得到 2 个重建的 A-pose 前景。计算其与输入的 A-pose 的前景的 *MSE* 并相加就能得到重建损失，数学定义见下式：

$$L = \sum_{r \in R} \left(\| \hat{C}_c(r) - C_c(r) \|_2^2 + \| \hat{C}_f(r) - C_c(r) \|_2^2 \right) \tag{7}$$

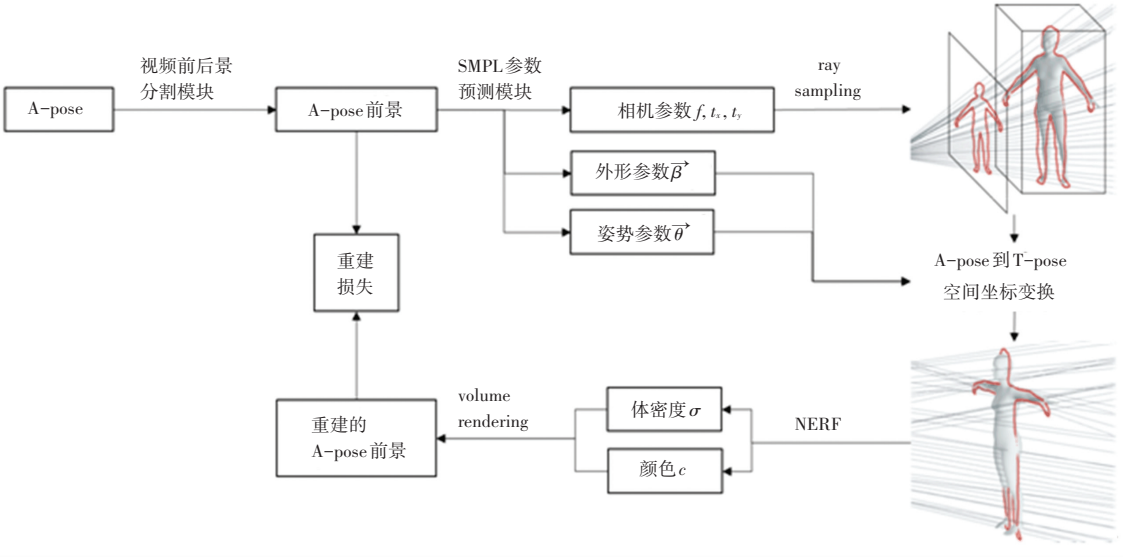


图 4 三维重建模块整体结构图

Fig. 4 Overview of the 3D reconstruction module

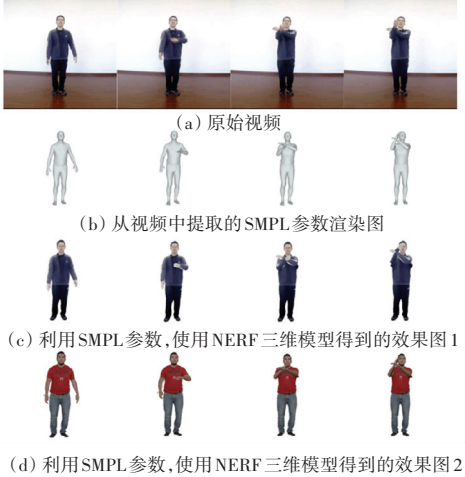
3 实验结果与分析

3.1 数据集

研究使用了 IPER[17] 数据集以及 people sanpshot[10] 数据集进行实验。people snapshot 包含 12 个人物,24 段视频。IPER 数据集包含 30 个人物,206 段视频,视频长度从 218 到 3 629 帧不等。平均每个视频的长度为 1 172 帧。

3.2 算法结果

图 5~图 7 是本文模型的效果图。图 5 中,第一行是原始视频,第二行是从视频中提取的 SMPL 参数渲染得到的,第三行和第四行则是根据 SMPL 参数,使用预训练的 NERF 三维模型重建得到的结果。图 6 修改了相机模型的参数,改变了重建视频的视角。图 7 则修改了人物的体型。



(a) 原始视频

(b) 从视频中提取的SMPL参数渲染图

(c) 利用SMPL参数,使用NERF三维模型得到的效果图1

(d) 利用SMPL参数,使用NERF三维模型得到的效果图2

图 5 算法效果图

Fig. 5 Results of the proposed method

(a) 原始视频

(b) 改变视角效果图1

(c) 改变视角效果图2

图 6 改变视角

Fig. 6 Change view

(a) 原始视频

(b) 改变体型效果图1

(c) 改变体型效果图2

图 7 改变体型

Fig. 7 Change shape

3.3 算法对比

效果对比如图 8 所示,文中将提出的模型与 Vid2Vid[1] 以及 FOMM[5] 算法进行了比较。Vid2Vid 和 FOMM 都是基于 GAN 的深度学习压缩方法,使用二维平面的运动信息配合关键帧进行重建。可以

看到当人体各个部分之间存在重叠的时候，二维平面的运动信息并不足以精确区分人体的各个部分。而本文所提的模型基于三维模型 SMPL，即使人体的各个部分存在重叠，依旧能够保证重建的效果。

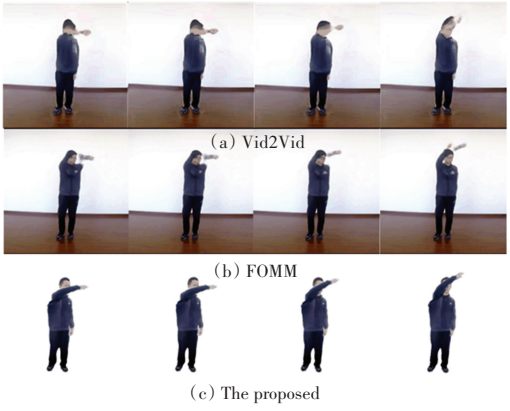


图 8　效果对比

Fig. 8　Comparison of the results

4　结束语

本文提出了一种基于 SMPL 的人物视频生成算法。首先，由人物 2D 图像重建出人物 3D 模型获取外观信息。然后，从视频中估计 SMPL 模型参数获取人物 3D 运动信息。最后，将人物外观信息与运动信息相结合生成人物视频。算法将人物视频生成从 2D 扩展到了 3D。实现多视角人物视频生成的同时，可以修改 SMPL 模型参数实现人物体型的改变。

参考文献

[1] WANG T C, LIU Mingyu, ZHU Junyan, et al. Video-to-video synthesis [J]. arXiv preprint arXiv:1808.06601 ,2018.

[2] MIRZA M, OSINDERO S J A P A. Conditional generative adversarial nets [J]. arXiv preprint arXiv:1411.1784, 2014.

[3] WANG T C, LIU Mingyu, TAO A, et al. Few-shot video-to-video synthesis [J]. arXiv preprint arXiv:1910.12713,2019.

[4] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. Animating arbitrary objects via deep motion transfer [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA:IEEE, 2019:2372-2381.

[5] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation [J].arXiv preprint arXiv:2003.00196v2,2020.

[6] 李园. 基于人体姿态迁移的视频生成方法研究 [D]. 西安：西安理工大学, 2021.

[7] 王红豫. 基于生成对抗网络的人体姿态合成人物图像与视频技术研究 [D]. 海口:海南大学, 2021.

[8] 张焜耀. 基于对抗网络的姿态迁移方法研究 [D]. 大连：大连理工大学, 2020.

[9] 赵宁, 刘立波. 融合自注意力机制的人物姿态迁移生成模型 [J]. 激光与光电子学进展, 2022, 59(04)：190-199.

[10] ALLDIECK T, MAGNOR M, XU W, et al. Video based reconstruction of 3D people models [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE, 2018:8387-8397 .

[11] 景成梁. 基于深度学习的三维重建与交互算法研究 [D]. 南京:东南大学, 2021.

[12] 张皓若. 基于 SMPL 模型的三维人体建模及其应用研究 [D]. 西安:陕西科技大学, 2020.

[13] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL：A skinned multi-person linear model [J]. ACM Transactions on Graphics,2015, 34(6)：1-16.

[14] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf：Representing scenes as neural radiance fields for view synthesis [C]//European Conference on Computer Vision. Cham:Springer, 2020：405-421.

[15] KOCABAS M, ATHANASIOU N, BLACK M J. Vibe：Video inference for human body pose and shape estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:1-14 .

[16] LIN Shanchuan, YANG Linjie, SALEEMI I, et al. Robust high-resolution video matting with temporal guidance [C]// 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).Waikoloa, HI, USA:IEEE, 2022:3132-3141.

[17] LIU Wen, PIAO Zhixin, TU Zhi, et al. Liquid warping GAN with attention：A unified framework for human image synthesis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence ,2024,44(9):5114-5132.

(上接第 70 页)

[6] SHAMOHAMMADI O,PAHLAVANI P, SHARIFI M. A. Comparison of the performance of gradient boosting, logistic regression, and linear support vector classifier algorithms in classifying travel modes based on GNSS data[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,2023,XLVIII-4/W2-2022.

[7] PODGORELEC V, KOKOL P, STIGLIC B, et al. Decision trees：an overview and their use in medicine.[J]. Journal of medical systems,2002,26(5):445-463.

[8] 邱伟栋. 基于 LightGBM 模型的 P2P 网贷违约预测研究[D]. 南昌:江西财经大学,2020.

[9] GUHA R A, UROLAGIN S. Credit risk assessment using decision tree and support vector machine based data analytics[C]// Proceedings of the 1st American University in the Emirates International Research Conference. Dubai:Springer International Publishing, 2019.

[10] 隋文涛,王文超,袁林,等. 基于 KNN 算法的铣刀状态监测技术研究[J/OL]. 机械设计与制造:1-4[2023-03-27].https://doi.org/10.19356/j.cnki.1001-3997.20230302.005.

[11] 杨涛,刘文杰,丁宁. 基于梯度下降算法的神经网络模型研究[J]. 网络安全技术与应用,2013(04):75-77.

[12] SCHAPIRE R, FREUND Y. Boosting [M]. Masssachusetts, USA：The MIT Press,2012.

[13] 崔佳旭,杨博.贝叶斯优化方法和应用综述[J]. 软件学报,2018,29(10):3068-3090.

[14] 孙斌,储芳芳,陈小惠. 基于贝叶斯优化 XGBoost 的无创血压预测方法[J]. 电子测量技术,2022,45(07):68-74.

[15] 李宁. 集成学习在银行个人信贷违约中的应用[D]. 重庆:西南大学,2022.