

足球运动员人物形象三维构建项目报告

代俊豪 王威 周旭东

2024 年 1 月 7 日

摘要

本项目旨在为普通人打造逼真、具备相应动作的足球运动员数字分身。通过采用用户提供的普通照片作为输入，经过 RoPE 模型进行二维换脸，再通过 TeCH 模型进行三维模型构建。在此基础上，通过 HybrIK 模型进行姿态识别，生成模型及其射门姿态系列图像。在方法层面，广泛尝试了不同模型的组合，包括 DeepFacelab、faceswap、RoPE、SHERF、InstantAvatar、Vid2Avatar、TeCH、ROPM、CLIFF、ProPose 和 HybrIK 等并进行了对比。在此过程中，本小组进行了多方面的改进和创新，包括对模型的适应性和灵活性提升，动画生成流畅性的改进，资源消耗的优化，以及姿态识别的精准和高效。通过这些创新和改进，本项目最终取得了显著的成果。

关键字：足球运动员数字分身，换脸技术，三维模型构建，姿态识别

Abstract

This project aims to create realistic and dynamically posed 3D digital avatars of soccer players for ordinary individuals. Utilizing regular photos provided by users, the process involves facial reenactment using the RoPE model for 2D facial mapping, followed by 3D model construction using the TeCH model. Subsequently, the HybrIK model is employed for pose recognition, generating a series of images capturing the avatar's shooting poses. Methodologically, a diverse set of model combinations, including DeepFaceLab, faceswap, RoPE, SHERF, InstantAvatar, Vid2Avatar, TeCH, ROPM, CLIFF, ProPose, and HybrIK, were extensively explored and compared. Throughout this exploration, the team implemented various enhancements and innovations, encompassing improved adaptability and flexibility of the models, enhanced animation smoothness, optimized resource utilization, and precise and efficient pose recognition. Through these innovations and improvements, the project achieved significant outcomes.

Keywords: Soccer Player Avatar, Facial Reenactment, 3D Model Construction, Pose Recognition

1 简介与意义/Introduction

1.1 项目意义和依据/Significance

随着科技的迅猛发展，元宇宙成为了人们关注的焦点，尤其在体育领域的运用更是引起了广泛的关注。在卡塔尔世界杯中，元宇宙技术的应用让体育赛事更加生动和真实，球迷们可以在虚拟空间中感受到身临其境的赛事情境。曼城与 Sony 联合推出元宇宙足球场的举措进一步强调了元宇宙技术在足球领域的巨大潜力，为球迷和普通大众提供了一种前所未有的沉浸式足球体验。为了进一步推动元宇宙技术在体育界的运用，本小组选择了足球员虚拟分身这个项目，以期在元宇宙中打造具有丰富身体结构、穿着和职业特点的足球运动员形象。

在中国，足球一直是备受瞩目的体育项目，但长期以来，中国足球一直面临发展的困境。尽管在基层足球普及方面取得了一些进展，但在职业足球水平上与国际先进水平仍有差距。元宇宙技术

的引入将为球员提供更为真实的训练和比赛体验，有助于提高球员的感知和反应能力。同时，通过虚拟分身的建立，球员们可以在元宇宙中更直观地了解自己的优势和不足，为自身的技术提升提供新的思路。

此外，通过元宇宙技术的应用可以创造更加引人入胜的足球体验，吸引更多普通大众参与，并为足球事业注入新的能量。普通大众可以在虚拟空间中感受真实的足球比赛。无需身临其境，球迷们可以通过虚拟分身在场边感受球场氛围、与其他球迷互动，甚至亲身经历球员的角度。这将激发更多人对足球的浓厚兴趣。

基于此，本项目致力于为普通人打造足球运动员数字分身，通过普通人照片输入，运用先进的图像处理技术和三维建模算法，以获得真实照片为基础，生成具有逼真外观和生动动作的足球运动员 3D 模型。这一模型将不仅限于静态的外观，更会包含精确的足球运动动作，其中射门动作将成为数字分身的一个独特之处。本小组将通过运用先进的计算机视觉技术捕捉并还原普通人的生动表情、体态和特征，同时结合足球运动员的专业动作，使数字分身呈现出真实感和足球专业性。

这项技术的应用不仅提供了一种独特的足球体验，也使得普通人可以在元宇宙中亲身感受到成为足球运动员的乐趣。通过个性化的数字分身，每个普通人都能在虚拟足球领域中找到属于自己的角色，与朋友互动，甚至参与虚拟足球比赛。这不仅为足球带来更为广泛的参与，也为元宇宙技术在娱乐领域的创新发展提供了一个引人注目的案例。

1.2 系统框架/Article Structure

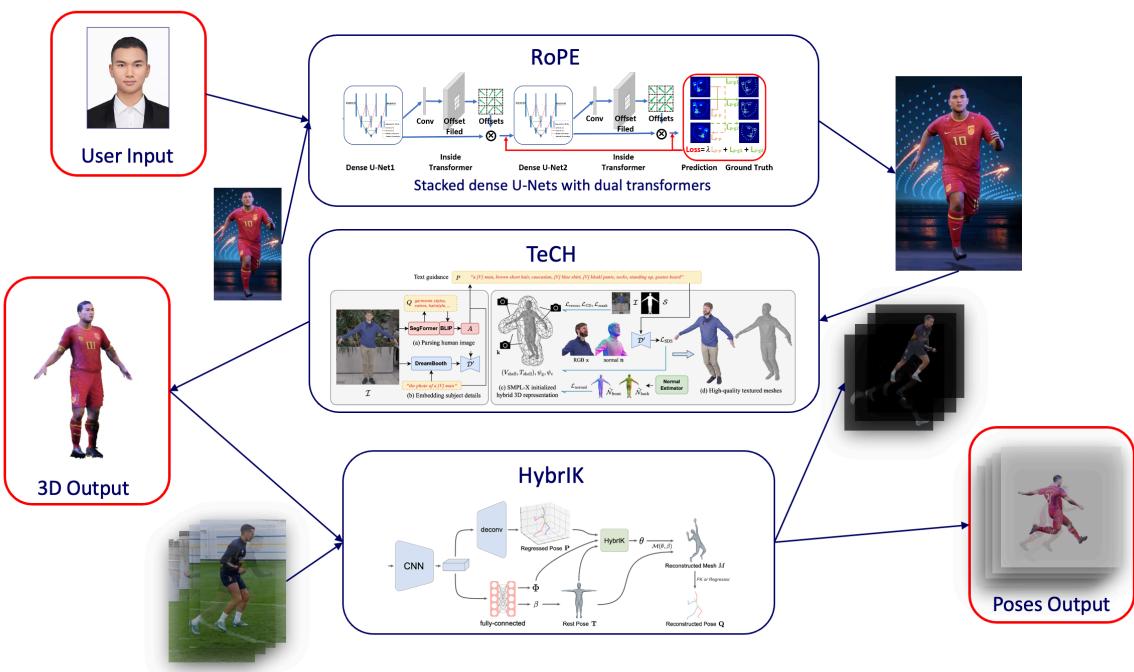


图 1: 系统总体框架

图 1展示了本项目的总体架构。项目以单张普通人脸（以小组成员代俊豪为例）作为输入，最终输出其足球运动员形象 3D 模型 (3D Output) 以及进行射门动作的一系列图像。在系统中引入一张足球运动员图 a 像与一系列踢球姿态图像 c 作为固定参考，中间能够生成换脸结果图像 b 和姿态参考图像 d。需要注意的是，这些“参考”不属于输入部分，可以更改任意输入图像 (User Input) 而

生成其对应面容的足球运动员模型及其射门动作场景。

系统整体流程为：获取输入图像 User Input，RoPE 对足球运动员参考图像 a 进行输入图像的二维换脸，获得换脸结果图像 b 后输入到 TeCH 模型中进行三维模型构建，得到 obj 模型输出（这里省略贴图部分）。在此基础上，通过参考姿态系列图像 c 通过 HybrIK 模型进行姿态识别后输出模型射门姿态系列图像，可以将这一系列图像合成为流畅的视频或动图。

2 技术路线选择/Techical route selection

实现从 2D 照片输入进行 3D 足球运动员构建有多条技术路线其一是对普通人照片进行三维重建后进行运动员衣服替换；二是将普通人面部进行三维构建后对运动员身体进行三维换脸；三是将普通人照片与运动员视频进行二维换脸后再进行统一三维构建。

第一条技术路线首先通过计算机视觉和深度学习技术，对输入的普通人照片进行三维重建。这涉及到对图像进行深度估计，从而获得照片中物体的三维结构。然后，使用三维模型的表面信息，系统可以准确地定位普通人的身体部位。接下来，通过足球运动员的 3D 模型，提取其衣服和身体结构信息。通过巧妙的图形处理算法，将普通人的三维模型与运动员的 3D 模型结合，实现对普通人穿着运动员衣服的替换。这一过程需要高级的图像合成技术，确保替换后的 3D 模型与原始照片相吻合，使得数字分身看起来自然且逼真。考虑到现有技术难以保证服装替换的效果，在该方案并未深究。

第二条技术路线着眼于人物面部的细节。通过先进的人脸识别和三维构建技术，系统可以对输入的普通人照片进行精确的三维面部重建，捕捉面部的特征。接着，从足球运动员的 3D 模型中提取身体结构信息。然后，将普通人的三维面部与运动员的身体结合，实现对普通人的三维换脸。这需要精准的面部迁移算法，以确保合成的数字分身在面部表情和特征上与原始照片相匹配，同时确保整体的逼真性。在这个方向上本项目组通过 gender-detection(1) 进行性别识别，3DDFA(2)(3)(4) 进行人物面部重建，取得初步成果。



图 2: 原图片及性别识别

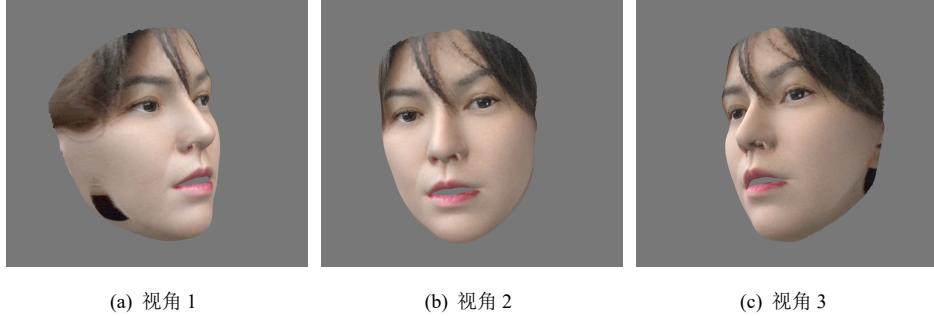


图 3: 脸部三维重建模型不同视角展示

以孙燕姿图像为例，性别识别能够获取准确结果如图 2(b)。同时，在图 3(a)3(b)3(c)中展示的人物脸部 3D 重建呈现出高度的精准度。然而，由于该技术是通过在预先定义的脸部上进行特征点对应贴图，脸型不匹配的问题较为明显；其次，人物发型的一部分包含在识别的面容中，无法很好和预设模型相匹配。在后续工作中，将脸部自动识别定位到身体相应部位而不出错也是困难的工作。最终，本小组选择第三条技术路线。

这一技术路线首先使用二维换脸技术，将普通人照片与足球运动员图片进行融合，使得普通人的面容在运动员身体上呈现。通过对照片与目标图片中的面部特征进行准确匹配和映射来实现。随后，通过对运动员图片中的运动姿势和身体结构进行三维重建，创建一个与运动员相匹配的 3D 模型，生成具有运动员身体结构和普通人面容的数字分身。在此基础上使用射门照片进行运动姿态识别，赋予数字分身射门动作并最终渲染导出相应效果。

3 相关工作/Related Works

本项目所选技术路线涉及三个重要过程，将分别阐述涉及的相关工作。

3.1 二维换脸/2D Face Swapping

在二维换脸领域，相关工作主要集中在利用深度学习技术进行面部特征的迁移和替换。经典的方法例如 DeepFaceLab(5) 和 FaceSwap(6) 利用生成对抗网络 (GAN) 模型，通过学习源图像和目标图像之间的映射关系，实现高质量的人脸替换。其中最为常见的是使用了 StyleGAN (Style Generative Adversarial Network) 等变种。它通过对抗训练的方式，同时训练生成器和判别器，使得生成器能够逐渐学到源图像到目标图像的映射关系。通过对大量数据的学习，DeepFaceLab 和 FaceSwap 能够生成高分辨率、逼真度很高的人脸替换结果。这些方法不仅适用于静态图像，还能在视频中实现动态的人脸替换效果。在近年来的研究中，RoPE(7) (Region of Perception Enhancement) 成为二维换脸技术领域的一项重要创新。RoPE 的目标是通过提高生成图像的感知质量，使得人脸替换的结果更为逼真和自然。该技术专注于处理生成图像中的细微细节和感知误差，以进一步提升人眼对于替换结果的真实感知。

3.2 三维模型生成/3D Model Generation

在三维模型生成领域，研究者们专注于通过图像或视频数据创建逼真的三维人脸模型。借助深度学习技术，尤其是自编码器等结构，能够从大量的图像数据中学得人脸的三维结构信息。NERF(8) 技术的引入为三维人物模型重建提供了强大的支持。NERF 采用基于神经辐射场的方法，通过训练深度神经网络来捕捉场景中的三维信息。这使得可以从单一或多个视角的图像数据中高精度还原物体表面及其细节，为人物模型的重建提供了高质量的输入。目前针对人物的 NERF 一般会以 SMPL(9; 10) 作为先验数据，以提升重建的效率。在此基础上，SHERF(11) 模型引入了稀疏的层次性编码，以更有效地处理大规模数据并提高模型的可扩展性。通过训练深度神经网络，该模型能够更精确地捕捉场景中的三维信息，从而在重建人物模型时提供更高质量的输入。

近期，InstantAvatar(12) 技术的应用进一步提升了三维人物模型的逼真性和交互性。该技术采用了快速、实时的人体姿态估计技术，能够在几秒内从单目视频中快速重建人类化身，并以交互速率实现动画和渲染。该系统采用了精心设计的神经场加速结构和高效的动态场景空白跳过策略，实现了高效的训练和优越的重建质量。其特点是训练时间（数十分钟）比常见的 NERF（数十小时）短很多。

此外，vid2Avatar(13) 通过直接在三维中建模场景中的人和背景，使用两个独立的神经场进行参数化，实现了场景分解和表面重建。并通过全局优化来处理背景模型、规范人体形状和纹理，以及

每帧的人体姿势参数。引入了一种由粗到细的体积渲染采样策略和新颖的目标，能够清晰分离动态人物和静态背景，实现了详细而稳健的3D人体几何重建。

3.3 模型姿态识别/3D Model Pose Recognition

在模型姿态识别的研究中，主要关注如何从二维或三维数据中提取并准确识别人物的姿态。本项目组尝试了最新技术训练姿态识别，其中，ROMP(14)(15)(16)同时预测身体中心热图和网格参数图，共同描述像素级别的3D身体网格。通过一个以身体中心为导向的采样过程，可以轻松从网格参数图中提取图像中所有人的身体网格参数。CLIFF(17)是一种采用对比学习思想的姿态估计方法，通过迭代反馈不断提升模型性能。在SHERF模型中便采用了CLIFF，显示该方法在其应用场景中取得了显著的效果。ProPose(18)则提出了一种新颖的分析公式，以贝叶斯方式学习基于骨方向的人体关节旋转的后验概率分布，该分布是在给定骨骼方向的条件下进行的。并基于此，提出了一个新的后验引导框架用于人体网格恢复。能够取得良好的效果。在工作中，HybrIK(19)(20)呈现出卓越效果，并被选定为最终识别器。

此外，OpenPose作为一款广受欢迎的开源工具包，在实时多人姿态估计方面表现出色。它能够检测和跟踪人体的关键点，涵盖了身体的各个部位，如头、手、肩膀等。InstantAvatar使用了OpenPose修正结果，涉及在姿态识别的基础上进行后处理或校正，以提高准确性。

这些相关工作为本项目的技术路线提供了有力的理论和实践基础，使得模型姿态得以准确识别，为数字分身的创造和应用提供更为可靠的基础。

4 研究内容与方法/Contnts and Methods

4.1 二维换脸/2D face swapping

4.1.1 DeepFaceLab

在2018年，DeepFakes(5)介绍了一个完整的生产流程，替换源人物的脸部并保持相同的面部表情，如眼睛运动和面部肌肉运动。然而，DeepFakes产生的结果在某种程度上表现不佳，Nirkin等人的自动人脸交换(21)也是如此。为了进一步唤醒人们对面部操纵视频的认识，并为伪造检测研究提供便利，开发者们建立了一个开源的Deepfake项目，DeepFaceLab（简称DFL）。

DeepFaceLab提供了一套形成灵活流程的工作流程。在DeepFaceLab中，可以将流程抽象为三个阶段：提取、训练和转换。此外DFL是一对一人脸交换模式，实现源(src)向目标(dst)的交换。

提取阶段：如图4所示，提取阶段是DFL的第一阶段，目的是从src和dst数据中提取人脸，该阶段包括多个处理部分例如人脸检测、人脸对齐、人脸分割。DFL提供了多种提取模式例如半脸(f)、全脸(wf)、整脸(head)，代表不同的人脸覆盖区域，本项目采用的是全脸模式。

- **人脸检测：**第一步是在给定的src和dst数据中找到目标人脸。DFL将S3FD作为其默认人脸检测器。也可以替换S3FD为其他人脸检测算法，如RetinaFace。
- **人脸对齐：**第二步是人脸的Alignment，DFL提供两种面部标志提取算法来解决：(a)基于热图的2DFAN（适用于标准姿势的脸）；(b)具有3D面部先验信息的PRNet（适用于具有大欧拉角（偏航、俯仰、滚动）的脸）。检测到面部标志后，还提供一个可选的函数，具有可配置的时间步长，用于在单个镜头的连续帧中平滑面部标志，以进一步确保稳定性。
- **人脸分割：**对齐后，得到了一个包含标准正面/侧视图的面部数据文件夹，在此基础上使用了细粒度的Face Segmentation网络(TernausNet)精确地对面部进行分割。由于在特定镜头中无法生成精细的遮罩，DFL引入了XSeg，允许使用少数的样本来训练模型。

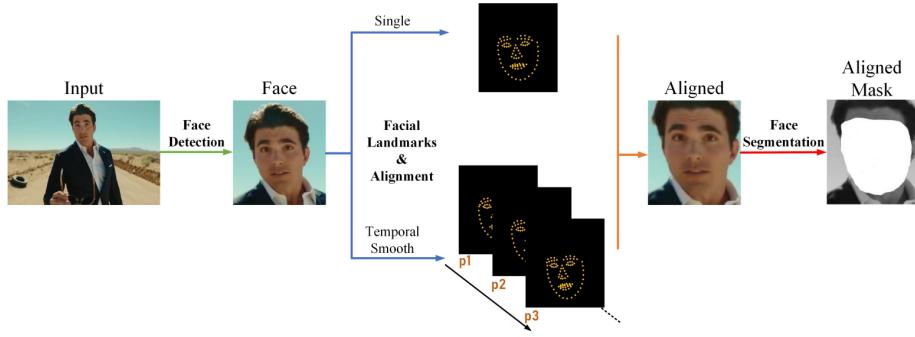


图 4: DeepFaceLab 提取阶段架构

训练阶段: 如图 5所示, DFL 提出了 DF 结构和 LIAE 结构来解决对齐的 src 和对齐的 dst 的面部表情不严格匹配同时保持生成的人脸的高保真度和感知质量的问题。

- **DF 结构:** 如图由 Encoder 和共享权重的 Inter 以及分别属于 src 和 dst 的两个 Decoder 组成。通过共享的编码器和 Inter 实现 src 和 dst 的泛化, 轻松解决了前述的不成对问题。DF 结构可以完成人脸交换, 但无法继承足够的来自 dst 的信息, 如光照。
- **LIAE 结构:** 为了进一步解决光照一致性问题, DFL 提出了 LIAE 结构, LIAE 是一个更复杂的结构, 具有共享权重的 Encoder、Decoder 和两个独立的 Inter。与 DF 相比, 主要区别在于使用 InterAB 生成 src 和 dst 的潜在代码, 而 InterB 仅输出 dst 的潜在代码。这里, $F^A B_{src}$ 表示源的潜在代码, 可将这个表示泛化为 F^B_{dst} 和 F^{AB}_{dst} 。在从 InterAB 和 InterB 获取所有潜在代码后, LIAE 通过通道融合这些特征图, $F^{AB}_{src} \parallel F^B_{src}$ 用于源的新潜在代码表示, 而 $F^{AB}_{src} \parallel F^B_{dst}$ 用于目标。然后, $F^{AB}_{src} \parallel F^B_{src}$ 和 $F^{AB}_{dst} \parallel F^B_{dst}$ 被输入到解码器中, 得到预测的源 (dst) 以及它们的掩码。串联 F^B_{dst} 是为了将潜在代码的方向转向需要的类别 (src 或 dst), 通过 InterAB 获得了源和目标在潜在空间中紧凑且 well-aligned 的表示。

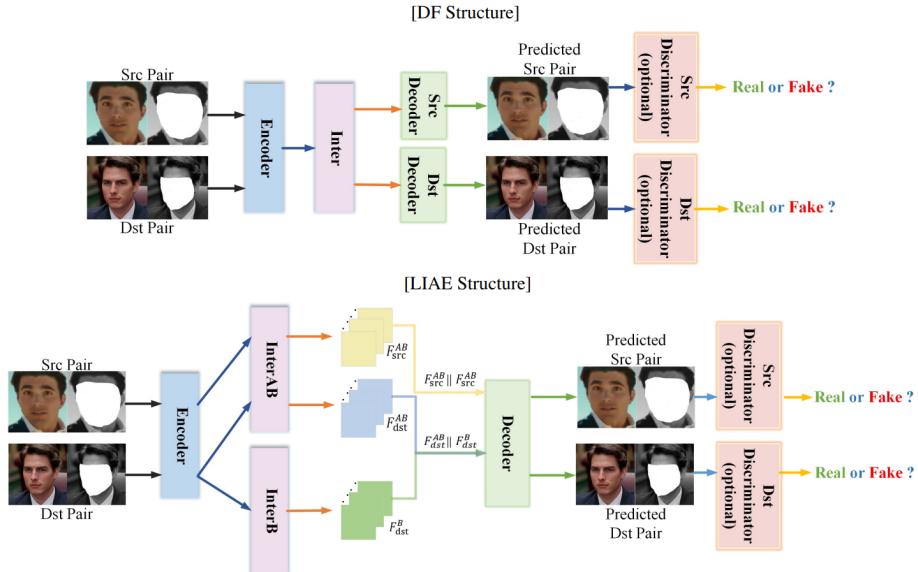


图 5: DeepFaceLab 训练阶段架构

转换阶段: 如图 6所示, 第一步是将从 dst 解码器生成的面孔及其遮罩从 dst 转换到 src 中目标

图像的原始位置 (22)。接下来是使重新调整的重现面孔与目标图像沿其外轮廓无缝地匹配。为了保持一致的肤色，DFL 提供了五种颜色转移算法（如 Reinhard 颜色转移：RCT (23)、迭代分布转移：IDT (24) 等），以将重现面孔的颜色近似到目标。任何混合都必须考虑到不同的肤色、面部形状和照明条件，特别是在重现面孔与被限定区域和目标面孔之间的交界处。DFL 通过 Poisson 混合 (25) 来实现这一点。

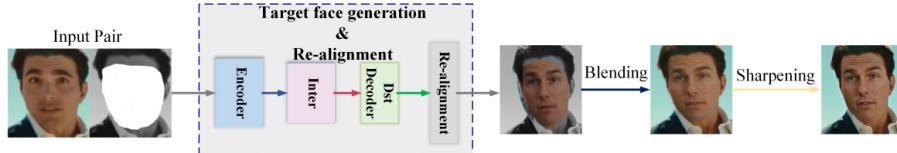


图 6: DeepFaceLab 转换阶段架构

但是 DeepFaceLab 也存在很多问题和不足。首先 DFL 是基于大量数据集的训练模型，在少量数据的情况下表现不佳，无法自行推断补充数据集，最终换脸效果差。其次 DFL 模型所需训练与迭代时间过长，按照估计，DFL 模型在数据集充足的情况下迭代约 100 万次能得到效果逼真的换脸结果。在本次实验条件下，使用少量数据集迭代 10000 次得到的结果在正面效果较好，在其他角度得到的效果较差。

4.1.2 Rope 模型

Rope 是 InsightFace 开源项目的 in_swapper model 的 GUI 集成应用。InsightFace 是一个面向人脸识别任务的开源项目，其目标是提供高性能、准确度高的人脸检测和识别技术，InsightFace 主要基于深度学习技术，特别是卷积神经网络 (CNN)。InsightFace 的核心功能包括人脸检测和人脸识别。人脸检测是指在图像中定位并标识出人脸的位置，而人脸识别则是将检测到的人脸与已知的人脸数据库进行匹配，以识别个体身份。InsightFace 不仅仅局限于人脸检测和识别，还涉及到其他与人脸相关的任务，如人脸属性分析、情感分析等。

InsightFace Alignment: Insightface 在人脸识别的准确度非常高，因此在后续的提取以及转换阶段表现更好。对于人脸图像的捕捉和面部标记，当前的技术水平主要围绕某些类型的深度卷积神经网络 (DCNNs)，如堆叠的 U-Net 和 Hourglass 网络。而 Insightface 创新性地提出了一种用于此任务的堆叠密集 U-Net，Insightface 设计了一种新颖的尺度聚合网络拓扑结构和通道聚合构建块，以提高模型的容量，同时不牺牲计算复杂性和模型大小。通过在堆叠的密集 U-Net 内使用可变形卷积和用于外部数据变换的一致性损失，模型获得了对任意输入面部图像的空间不变性的能力。

当前 2D 面部 Alignment 技术的主流，包括 U-Net 和 Hourglass 等模型，它们采用对称结构和多步池化以捕获不同尺度的局部和全局特征，并通过跳跃连接保留空间信息。为了提高模型的容量，Insightface 引入了深度层聚合 (DLA) 和提出了尺度聚合拓扑结构 (SAT)。然而，由于计算复杂性和模型大小的增加，SAT 在模型训练时遇到了优化困难。因此，Insightface 通过简化 SAT 结构，包括减少一个池化步骤和改变一些卷积方式，保持了与 Hourglass 相似的计算复杂性和模型大小，但显著提高了模型的容量。其次，为了增强模型对输入面部图像的空间不变性，Insightface 在堆叠的密集 U-Net 中引入了变形卷积和连贯损失的概念。这些变动允许模型更好地适应输入图像中的几何变换，并通过在模型外部应用一致的损失来确保模型的预测与不同变换一致。

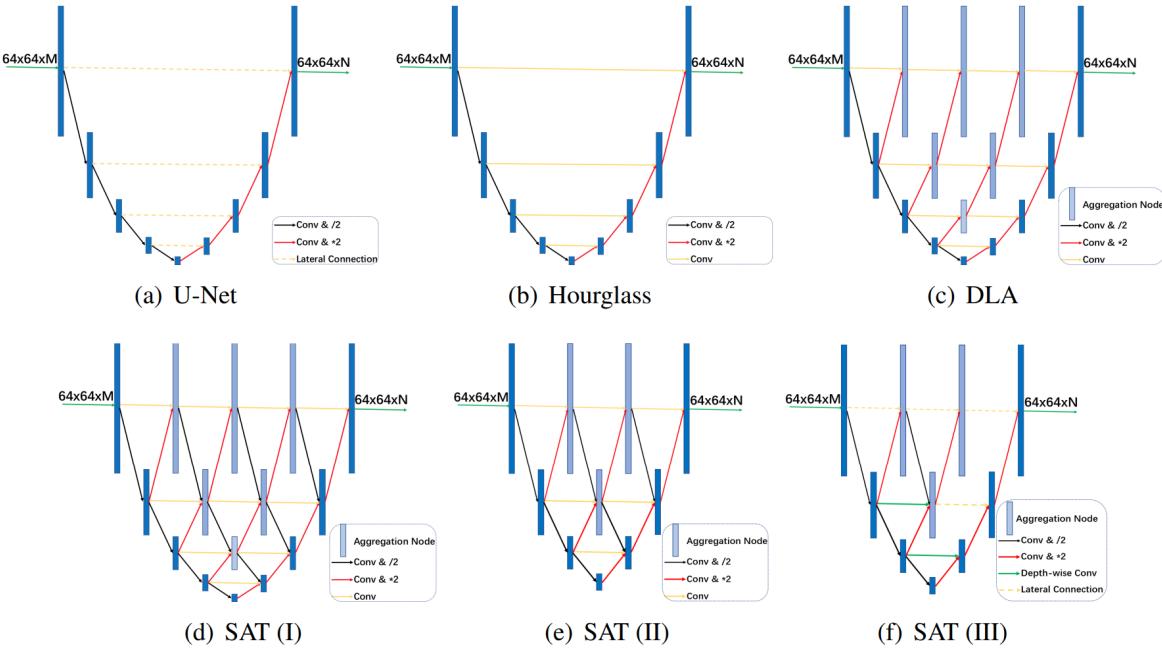


图 7: 不同的网络拓扑结构

原始的 Hourglass(26) 使用瓶颈残差块 (图 7(a))。为了提高块的容量, (27) 中探讨了一个并行和多尺度的 Inception 残差块 (图 7(b))。同时, 在(28; 29) 中广泛研究了一种新颖的分层、并行和多尺度 (HPM) 残差块 (图 7(c))。对于建模模块设计, 遵循网络拓扑中的相同见解, 并创新性地提出了一个通道聚合块 (CAB)。如图 7(d) 所示, CAB 在通道上对称, 而 SAT 在尺度上对称。输入信号在每个通道减少之前分支出, 并在每个通道增加之前汇聚, 以保持通道信息。在骨干部分进行的通道压缩有助于上下文建模(30), 其中包括通道级的热图关系, 并在局部观察模糊时提高鲁棒性。为了控制计算复杂性并压缩模型大小, 在 CAB 内部使用了深度可分离卷积(31)和基于复制的通道扩展。

Insightface 通过将两个 U-Net(26; 28) 端对端堆叠来增强模型的能力, 将第一个 U-Net 的输出作为第二个 U-Net 的输入。堆叠 U-Net 提供了自下而上、自上而下的推断机制, 可重复评估局部热图预测和全局空间配置。然而, 由于固定的几何结构, 堆叠的 U-Net 在变换建模方面仍有局限。为了解决这个问题, Insightface 考虑了两种不同类型的空间变换器: 使用 STN(32) 执行的参数显式变换和使用可变形卷积(33)执行的参数隐式变换。通过在第一个 U-Net 后应用 STN, 消除了刚性变换对输入面部图像的影响, 使得后续的 U-Net 只需关注非刚性面部变换。同样, 使用可变形卷积也达到了相似的效果。Insightface 选择可变形卷积作为内部变换器, 因为它更灵活地建模几何变换, 并具有更高的计算效率。

此外, 相比于 DeepFaceLab, in_swapper 模型可以实现单张输入图片提取人脸并换脸到目标图片或视频上, 因此在便捷性上, in_swapper 模型更胜一筹。实验测试得到结果 in_swapper 换脸模型在面部识别、细节调整、渲染速度等方面的确优于 DeepFaceLab, 对比如图 8 所示。

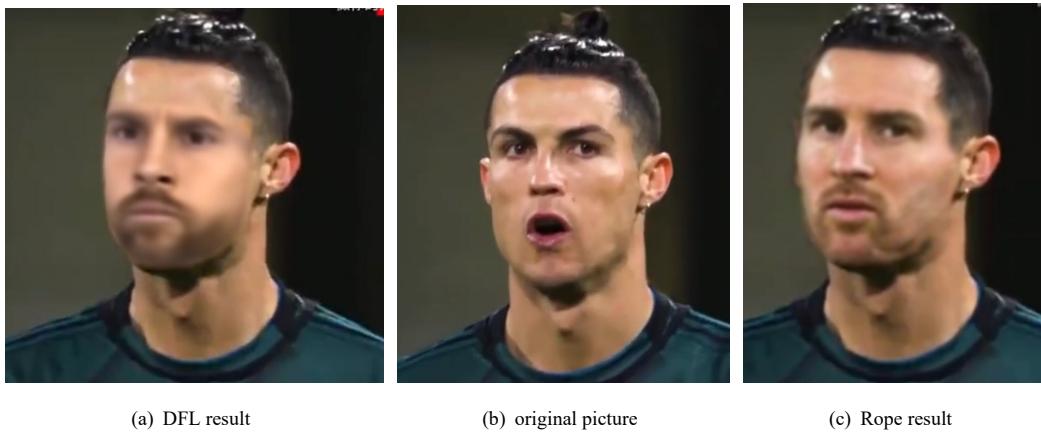


图 8: DFL 模型与 Rope 模型换脸结果对比

4.2 三维模型生成/3D model generation

4.2.1 SHERF

SHERF(11) 属于 NeRF 范畴，目标是从 2D 观察中建立高质量的人体。SHERF 引入了一个包括全局、点级和像素对齐特征的分层特征库，以解决单一图像输入中缺失的信息问题。分层特征库捕捉了全局人体结构和局部细节，这对于高保真度的人类 NeRF 重建至关重要。引入了一个特征融合 Transformer，以有效地合并分层特征库中的特征。SHERF 将 3D 人体表示建模为规范空间，使其易于进行姿态变换和渲染。SHERF 是第一个能够从单一人体图像中恢复可动画的 3D 人体的可推广 NeRF 模型，扩展了在实际场景中应用人类 NeRF 的可能性。

SHERF 不需要重新训练，可以使用预训练模型进行推理，达到不错的效果。

创新点：然而，SHERF 的开源代码是为科研准备的，仅考虑了数据集的输入与评估，不能实现端到端的模型生成。本项目将 SHERF 和 CLIFF(17) 结合，根据 CLIFF 的输出结果编写了数据集加载器，实现了单张图片作为重建输入、视频作为姿态输入，得到重建人物动画的流程。

4.2.2 InstantAvatar

InstantAvatar(12) 使用视频作为输入。但实际上，视频是被切分为单帧图片输入的，从这个意义上，其输入是大量的图片，为其提供了大量信息，但不包括帧之间的联系信息。InstantAvatar 通过提出一种方法，可以在不超过输入视频捕获时间的情况下，在 60 秒内重建高保真度的化身，迈出了向单目神经化身重建在实际应用中更具可行性的重要一步。为了实现这种加速，文章使用了 Instant-NGP(34) 和 Fast-SNARF(35) 两个组件。Instant-NGP 利用最近提出的高效神经辐射场变体来学习规范形状和外观。Fast-SNARF 是一种高效的关节模块，使学习能够从姿态观察中进行，并能够为人体添加动画。

与 SHERF 不同, InstantAvatar 和 Vid2Avatar 需要经过训练, 这更符合传统意义上的 NERF 特点。精准的 SMPL 估计作为先验数据, 对模型生成而言至关重要, 这一点在 SHERF 和 InstantAvatar 中是一样的。在 SHERF 的测试中, 项目组观察到了 CLIFF 的不精确导致的结果偏差。由于 ROMP 技术较老, 识别效果与 CLIFF 还有一定差距。对此, InstantAvatar 在数据预处理中使用了 OpenPose 修正 SMPL 模型, 其预处理流水线应是从 Neuman(36) 继承而来。这使 InstantAvatar 在保证较短训练时间的同时, 获得了更加精确的效果。

创新点: 在最后的动作生成阶段,项目组发现CLIFF的结果会导致帧间抖动,动作不流畅,于是本项目使用了更加准确的ProPose(18)进行姿态识别、动作提取,使最终动作不再有“卡顿”的感觉,更加流畅。

4.2.3 TeCH

尽管最近在从单一图像中重建穿着衣物的人类方面取得了研究进展，但精确还原具有高级细节的“未见区域”仍然是一个未解决的挑战，缺乏足够的关注。有的工作试图基于可见的视觉线索（例如颜色、法线）预测不可见区域（例如背部），但这通常导致模糊的纹理和平滑的几何形状，而且从不同角度观察模型时，会出现不一致性。解决这个问题的一个做法是引入多视图监督。但在只有单一图像的情况下，这个问题有没有可能解决呢？TeCH 与之前的工作不同，使用生成式模型来指导重建过程。提到图像生成，人们常会想到著名的 Stable Diffusion(37)，但它并不能对单一物体生成具有一致性的图片，因此需要使用 Dreambooth(38) 进行微调。其整体架构如图 9。

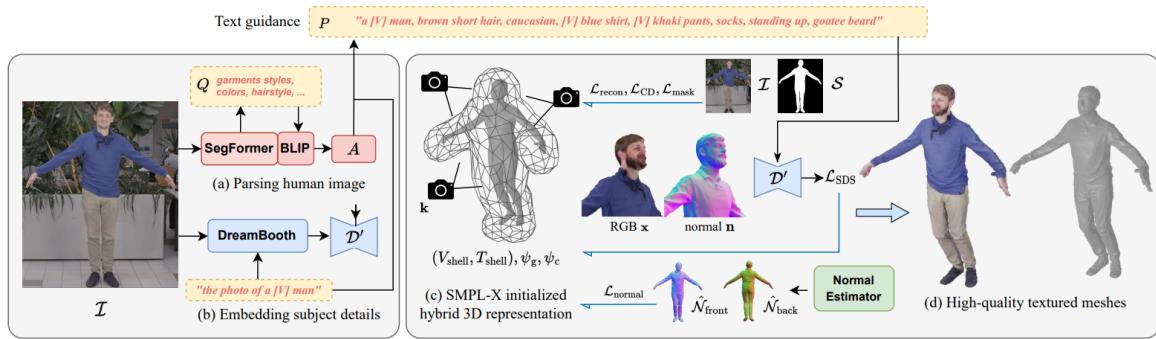


图 9: TeCH 架构

总的来说 TeCH 的工作流程分为以下几个阶段：

提取人体特征 从单一图像中获取的信息可以分为两类，如图 10(a)、(b):

- 可描述信息** 一个人的性别、头发颜色、衣服款式、面部特征等，是可以显性描述出来的；
- 不可描述信息** 然而，还有很多信息是难以直接描述的，比如衣服上的图案，人的高矮、胖瘦等，他们暗含了主体独特的特征。

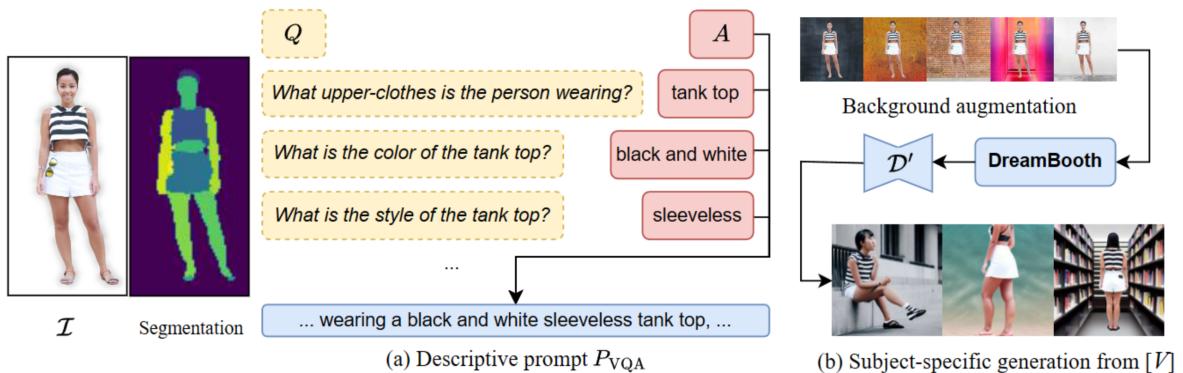


图 10: 提取人体特征

对于可描述的信息，可以直接使用 Prompt 描述。对于给定的单一图像输入，首先使用人体解析模型 SegFormer⁽³⁹⁾将人划分为帽子、裙子、裤子、鞋子等不同部分，然后使用视觉问答（Visual Question Answering, VQA）模型 BLIP⁽⁴⁰⁾对人体的不同部分提问，得到一系列问题的答案。这些问题涵盖了服装款式、颜色、面部特征、发型等，对人体的可描述信息做了全面的总结。随后，问题及答案被插入预定义的模板中，成为 P_{VQA} ，即所需 Prompt。

对于不可描述的信息，使用 Dreambooth 来捕捉。Dreambooth 是对扩散模型微调的一种方法，仅使用几张图片就可以得到个性化的模型。TeCH 在预训练的 Stable Diffusion 1.5 上进行 Dreambooth 微调。微调过程中有以下几点比较特别：

- 为了让模型记住人物特征，需要一个独特的标记词，TeCH 选用了“sks”，这虽然是一种枪的名字，但不常见；
- 一张图片数目太少，容易使模型记住背景，这对结果是不利的，因此 TeCH 将同一个人物复制到 5 张不同的背景上，共同作为输入；
- 微调 Dreambooth 需要“正则化图像”来避免模型产生语义漂移，TeCH 使用 VQA 判断人物的性别，将“man”和“woman”作为分类词，各使用 200 张图片作为正则化图像。

两种信息相辅相成才能达到最佳效果，单独使用哪一种都会导致不可见区域生成效果差。在后文的重建过程中，扩散模型是重要的一环。扩散模型可以生成各个方向的图片。观察到这些图片中人物脸部一般不清晰，TeCH 又让扩散模型贴近人物的脸部，生成脸部细节大图。

创新点：原论文中使用的 Dreambooth on Stable Diffusion¹，其微调过程消耗显存极大，以至于作者直接写出最小显存要求是 $2 \times 32G$ ，这对大多数人来说是极为苛刻的条件。本项目改进工作流程，采用了在扩散模型上运行的 Dreambooth 另一实现²，这一实现可以启用混合精度、8 位 Adam 优化器等特性减小显存占用。根据其介绍，最低可以在仅有 8GB VRAM 的 GPU 上实现微调。在 Dreambooth 的微调阶段，项目组发现如果训练步数过多，会使扩散模型过拟合，不论 Prompt 如何，都会输出固定的结果，无法生成不可见区域；如果步数过少，则难以形成记忆，输出结果与微调内容不相符。经过反复实验，本项目采用 300 步（在论文采用的模型上是 800 步），在这个数值下，模型可以输出不同角度的图片（通过在 Prompt 中加入“back view”“the face of”等词，如图 11），而又不至于与原输入差距过大。



图 11：扩散模型微调后的生成结果

¹<https://github.com/XavierXiao/Dreambooth-Stable-Diffusion>

²<https://github.com/huggingface/diffusers/tree/main/examples/dreambooth>

几何重建 为了更好地在高分辨率下表现 3D 人物，TeCH 采用了混合的几何表示，同时使用了 SMPL-X(10) 和 DMTet(41) 两种模型，共同表示人物的几何外观。首先通过 PIXIE(42) 估计一个初始身体，再对初始身体应用网格扩展、下采样、上采样等一系列操作，以获得外围包裹身体的“壳”。

在几何重建阶段，损失函数包括剪影损失（基于原始图像）、文本指导 SDS 损失（基于微调过的个性化扩散模型）、几何正规化（基于法线与光滑度）。TeCH 在 DMTet 上执行从粗到细的划分，以更稳定地为穿着衣服的人生成高分辨率的网格。

最后，通过 Marching Tetrahedra(43) 从四面体网格中提取所需网格。在进入下一阶段前，将 SMPL-X 与网格对齐，这样就可以通过 SMPL-X 为模型提供动作。此外，由于这一阶段对手的生成通常不理想，TeCH 还将原网格中的手去掉，替换为 SMPL-X 中的手，方便下一阶段的纹理重建，也提升了整体的效果。

纹理重建 为了恢复一致的细节与颜色，TeCH 在此阶段使用两个姿态进行渲染及优化：输入姿势和 A-Pose。初始时纹理为完全随机生成。

在纹理重建阶段，损失函数包括遮挡感知的重建损失、文本指导 SDS 损失、颜色一致性损失。其中，遮挡感知的重建损失是指，将重建损失直接应用在自遮挡区域可能由于几何错位而导致不正确的纹理，因此 TeCH 使用掩码排除了这一部分。

动作生成（创新点） 这一部分在原论文中没有详细介绍，本项目参考 ECON(44)，探索出了将 SMPL-X 动作应用到 TeCH 生成结果上的方法。项目组观察到，在几何重建与纹理重建之间，TeCH 将模型网格与 SMPL-X 对齐，并生成了 A-pose、大字形等姿态。查看这一部分的代码，可以发现 TeCH 有能力将网格转换为 SMPL-X 参数所代表的姿态。但是由于此时还没有生成纹理，直接导出网格并动画是不可行的。尝试过绑定骨骼后，项目组发现绑定难度大，产生动作不自然，且忽视了模型中已经存在的 SMPL-X 特征。

研读了 ECON 的相关代码后，项目组找出了生成带纹理动画的方法。获取最后网格结果后，可以将其以 .obj 形式导入建模软件 Blender 中，可以得到有纹理、无动作的静态模型；在几何重建后，可以得到无纹理、SMPL-X 姿态的静态模型。因此，本项目编写了两个脚本来实现动画生成：

- 在几何重建后，输入由 HybrIK 估计的 SMPL-X 参数，产生一系列（由视频帧数决定）的顶点位置，将这些顶点位置保存备用。
- 在 .obj 导入 Blender 时，需要选择“保持顶点顺序”，得到有纹理的模型。脚本通过逐帧将模型顶点位置替换为上述位置，以实现模型的运动，同时保持 UV 与材质。暂时没有找到直接导入 UV 的办法，因此还是采用先导入带纹理模型、再替换顶点坐标的方式。此处涉及到不同坐标系的坐标换算。随后便可以使用 Blender 渲染得到结果，详情在实验结果与分析中。

4.3 模型姿态识别/3D Model Pose Recognition

4.3.1 ProPose

Propose 是一款专注于人体网格恢复任务的深度学习框架，通过学习概率分布来建模人体姿态的不确定性，特别关注于解决关节旋转的不确定性。采用新颖的基于矩阵 Fisher 分布的学习友好和数学上正确的概率分布形式，以贝叶斯方式回归 3D 关节旋转的后验概率分布，理论上证明后验概率更加集中，有助于神经网络的学习。除了理论贡献外，Propose 提出了一个新的人体网格恢复框架，成功实现了高精度和高鲁棒性，并在性能上优于现有基准方法。

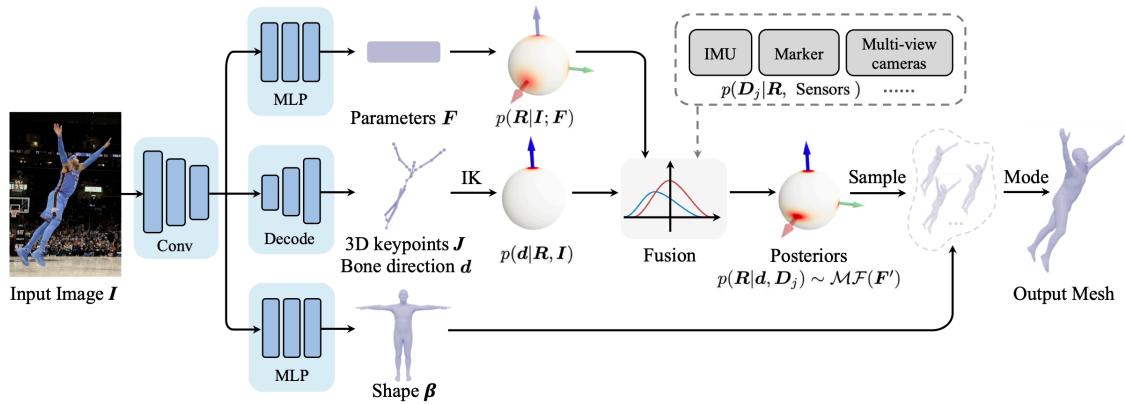


图 12: ProPose 架构

如图 12 所示, ProPose 系统接受输入图像并通过多分支网络同时预测先验矩阵 Fisher 参数 F 、3D 关键点 J 以及 SMPL 形状参数 β 。通过从 3D 关键点 J 计算骨骼方向 d , 作为在 3D 旋转条件下的似然度。使用贝叶斯规则进行后验概率计算 (Fusion), 该概率仍然遵循矩阵 Fisher 分布, 但具有不同参数和更大置信度, 同时可融合来自其他传感器的附加观测数据。最终, 利用估计的旋转和形状参数, ProPose 框架能够还原人体网格, 实现准确的人体姿态估计。

该识别器的主要贡献点有:

- 新颖的概率分布学习公式: Propose 提出了一种新颖、学习友好且数学上正确的概率分布学习公式, 通过矩阵 Fisher 分布建模旋转不确定性。这一公式允许以贝叶斯方式回归 3D 关节旋转的后验概率分布, 并在解析形式上进行。
- 新的人体网格恢复框架: 引入了一个新的人体网格恢复框架, 利用学到的解析后验概率。该框架实现了在相同时间达到高精度和高鲁棒性的目标, 并在性能上优于现有的基准方法。
- 灵活的多传感器融合机制: 提出了一种新颖且灵活的多传感器融合机制, 使得框架可以与额外的传感器 (如多视角摄像头、光学标记、惯性测量单元) 无缝集成。与传统的多传感器融合算法不同, Propose 的框架允许在训练阶段进行融合, 从而更好地学习传感器的噪声特性, 并有潜力提供更高的精度。

4.3.2 HybrIK

在 TeCH 模型生成的基础上, 本项目组使用 HybrIK 姿态识别器。这是一种混合分析-神经逆运动学解决方案, 其中逆运动学用于从 3D 身体关节中找到相应的身体部分旋转, 旨在从视觉内容中恢复完整的三维人体表面。该方法使用混合分析-神经逆运动学算法, 通过 3D 关键点估计和整体身体网格估计之间的协作关系, 提高了图像-网格对齐的准确性。

传统的优化方法和学习方法在解决这一问题上存在挑战。该研究创新性地利用了 3D 关键点估计的思想，通过建立 3D 关键点和身体网格之间的联系，解决了图像-网格不对齐的问题。HybrIK 使用逆运动学 (IK) 来从 3D 关节中找到相应的身体部分旋转，通过创新性的扭转和摆动分解解决了这一非唯一解问题。

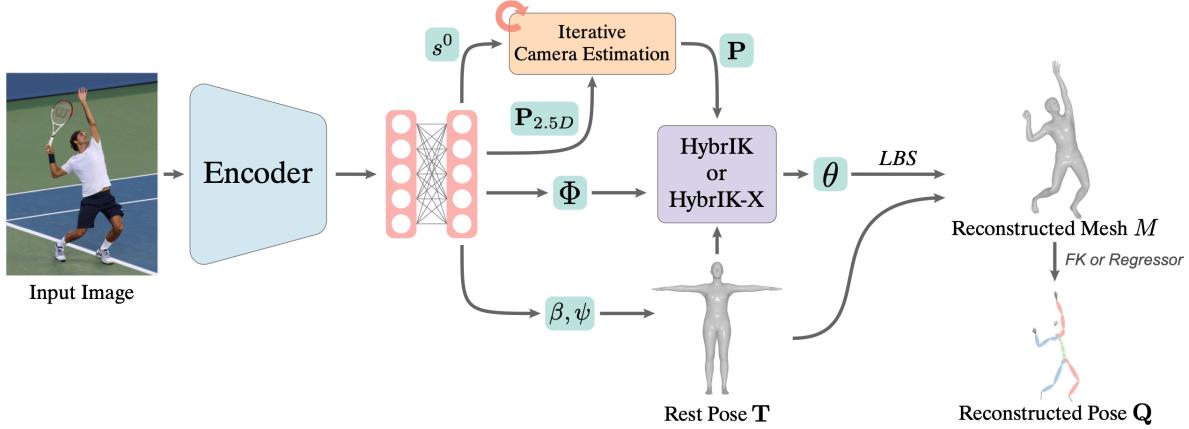


图 13: HybrIK 架构

如图 13 所示，HybrIK 系统通过神经网络学习的 2.5D 关节、形状参数、表情参数、扭转角度和初始摄像机参数是从视觉线索中获取的。这些学到的 2.5D 关节和初始摄像机参数被输入到迭代摄像机估计算法，产生估计的摄像机和反投影的 3D 关节。这些结果传递到 HybrIK 过程，解决身体部分旋转，最终通过姿势和形状参数获得重建的身体网格和姿势。

该识别器的主要贡献点有：

- 提出了一种新颖的整体身体网格恢复框架，使用混合分析-神经逆运动学算法将准确的 3D 关节转化为像素对齐的身体网格。
- 通过在 3D 骨架和参数模型之间建立联系，提高了身体网格恢复的图像-网格对齐，并同时解决了 3D 关键点估计方法中不切实际的身体结构问题。
- 在各种仅身体、仅手和整体身体基准测试中取得了最先进的性能。

5 实验结果与分析/Results and Analysis

5.1 效果展示/Result Demonstration

由 TeCH 进行三维人物构建效果如图 14 所示，可以看到人物面容，服装和体态等特征清晰，面容尤为清晰，这是要求扩散模型贴近“观察”的结果；尤其值得注意的是，在仅有单张图片作为输入的情况下，该结果显得格外出色。但是由于扩散模型的特征，其部分构建（特别是不可视区域）准确度不足，观察到胸前存在图案重叠错乱，背后出现数字混乱等情况。但总体而言，特征显著，瑕疵并未掩盖其优越之处。



(a) 视角 1

(b) 视角 2

(c) 视角 3

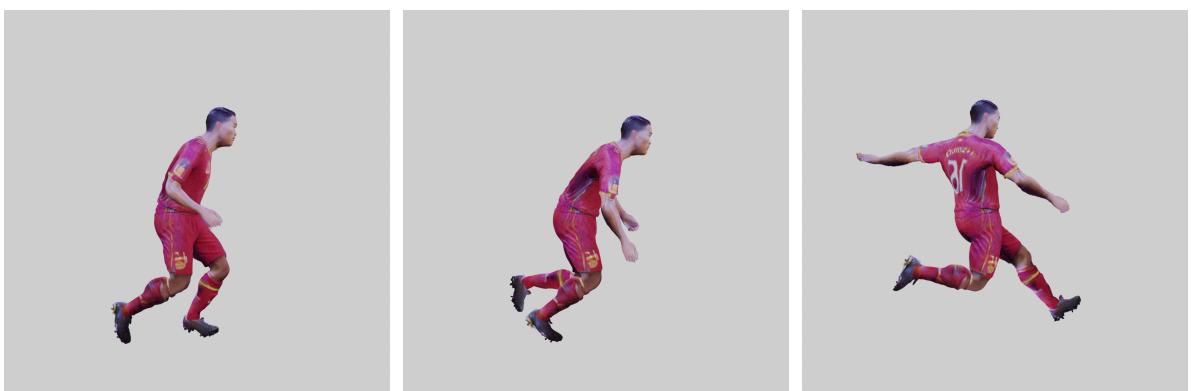
图 14: TeCH 模型生成效果



(a) 参考姿态 1

(b) 参考姿态 2

(c) 参考姿态 3



(d) 识别姿态 1

(e) 识别姿态 2

(f) 识别姿态 3

图 15: 参考过程姿态及 HybrIK 识别结果

从图 15 中可以看出，HybrIK 对于模型骨架解析十分准确，姿态识别效果出色。

5.2 实验结果对比/Comparison of results



图 16: 模型构建结果对比

从图 16 的对比中可以看出, SHERF 有严重依赖正面图像, 对其他方向的构建十分粗糙。InstantAvatar 能够给出较为精确的人物模型, 但面部复制等模糊程度高, 并带有空间中的噪声。而 TeCH 的结果显著地优于前两者。



图 17: 模型姿态识别结果对比

从图 17 可以看出, 两个模型对动作识别都十分精确, Propose 在姿态表现上较为柔和, 而 HybrIK 在躯干呈现上更加清晰。此外, HybrIK 对 TeCH 所使用的 SMPL-X 模型, 在通用性上有更好的表现。

6 特色与创新/ Distinctive or Innovation Points

在宏观层面上, 本项目将换脸技术与重建技术巧妙结合, 从整体上实现了一种更为综合的方法。从微观层面来看, 本项目进行了以下几点改进和创新:

- **SHERF:** 只支持数据集形式的输入，通过为 CLIFF 编写数据加载器，实现了对数据集形式输入的支持，从而提升了系统的灵活性和适用性。
- **InstantAvatar:** 在动画生成方面采用了 ProPose，使动画呈现更加流畅，提高了用户体验。
- **TeCH:** 将论文原模型替换为显存消耗更小的 DreamBooth 模型，使其运行的最低要求从 $2 \times 32G$ 降低到单卡。这一创新降低了系统的资源消耗，保持了性能水平。同时，调整了 Dreambooth 模型的微调步数，取得了较为理想的结果。最终，TeCH 在 RTX 4090 上需要大约 5 小时时间训练。
- 在数据预处理阶段的生成 Mask 阶段，采用了在 InstantAvatar 中使用的更优秀的 SAM 替代，取得了更好的效果。
- 使用支持 SMPL-X 的 HybrIK-X 生成动作。
- 参考 ECON，为 TeCH 编写动作生成器与 Blender 渲染脚本。

创新具体内容在研究内容与方法一章中有详述。

7 总结/Conclusion

本项目的目标是打造一个创新性的数字分身系统，使普通人能够享受逼真且具有个性的足球运动员体验。通过用户提供的一般照片，设计并整合了 RoPE、TeCH 和 HybrIK 等先进模型，构建了一个复杂而高效的流程，从而实现了在静态图像中生成生动且动作丰富的三维足球运动员模型的目标。

在方法层面，本小组进行了大量的模型组合尝试，涉及 DeepFacelab、faceswap、SHERF、InstantAvatar、Vid2Avatar、ROPM、CLIFF 等，并与之前提到的模型进行对比。这些组合实验的结果为选择最优方案提供了深刻的见解。本小组不仅着眼于提升模型的适应性和灵活性，更专注于改善动画生成的流畅性，优化系统资源的消耗，以及提高姿态识别的准确性和效率，取得了一系列显著的创新成果，如 SHERF 模型的数据输入支持、InstantAvatar 动画生成的 ProPose 优化、TeCH 模型的资源消耗降低，以及 SAM 的替代使用等。

最终，该数字分身系统不仅在逼真性和动作表现上取得了令人满意的效果，而且在系统的整体性能上也达到了令人瞩目的水平。通过对多个模型和方法的深入比较和细致改进，本项目为数字人体建模和动作生成领域的研究贡献了有价值的经验和成果。

参考文献

- [1] A. Ponnusamy, “gender-detection-keras,” 8 2019. [Online]. Available: <https://github.com/arunponnusamy/gender-detection-keras>
- [2] J. Guo, X. Zhu, and Z. Lei, “3ddfa,” <https://github.com/cleardusk/3DDFA>, 2018.
- [3] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, “Towards fast, accurate and stable 3d dense face alignment,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [4] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, “Face alignment in full pose range: A 3d total solution,” *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [5] iperov, “Deepfacelab,” <https://github.com/iperov/DeepFaceLab>, 2020.
- [6] deepfakes, “faceswap,” <https://github.com/deepfakes/faceswap/>, 2020.
- [7] Hillobar, “Rope,” <https://github.com/Hillobar/Rope>, 2023.
- [8] M. T. J. T. B. R. R. N. Ben Mildenhall, Pratul P. Srinivasan, “Nerf: Representing scenes as neural radiance fields for view synthesis,” *ECCV Computer Vision and Pattern Recognition (cs.CV); Graphics (cs.GR)*, 2020.
- [9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
- [10] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolckart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body capture: 3d hands, face, and body from a single image,” in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [11] S. Hu, F. Hong, L. Pan, H. Mei, L. Yang, and Z. Liu, “Sherf: Generalizable human nerf from a single image,” *arXiv preprint arXiv:2303.12791*, 2023.
- [12] J. S. O. H. Tianjian Jiang, Xu Chen, “Instantavatar: Learning avatars from monocular video in 60 seconds,” 2020.
- [13] X. C. J. S. O. H. Chen Guo, Tianjian Jiang, “Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition,” 2023.
- [14] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, “TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments,” in *CVPR*, 2023.
- [15] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, “Putting People in their Place: Monocular Regression of 3D People in Depth,” in *CVPR*, 2022.
- [16] Y. Sun, Q. Bao, W. Liu, Y. Fu, B. Michael J., and T. Mei, “Monocular, One-stage, Regression of Multiple 3D People,” in *ICCV*, 2021.
- [17] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, “Cliff: Carrying location information in full frames into human pose and shape estimation,” in *ECCV*, 2022.
- [18] Q. Fang, K. Chen, Y. Fan, Q. Shuai, J. Li, and W. Zhang, “Learning analytical posterior probability for human mesh recovery,” in *CVPR*, 2023.

- [19] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, “Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383–3393.
- [20] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, and C. Lu, “Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery,” *arXiv preprint arXiv:2304.05690*, 2023.
- [21] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, “On face segmentation, face swapping, and face perception,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 98–105.
- [22] S. Umeyama, “Least-squares estimation of transformation parameters between two point patterns,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.
- [23] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, “Color transfer between images,” *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.
- [24] F. Pitié, A. C. Kokaram, and R. Dahyot, “Automated colour grading using colour distribution transfer,” *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 123–137, 2007.
- [25] P. Perez, ““poisson image editing” by perez, blake and gangnet.”
- [26] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 483–499.
- [27] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou, “Cascade multi-view hourglass model for robust 3d face alignment,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 399–403.
- [28] A. Bulat and G. Tzimiropoulos, “Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3706–3714.
- [29] ——, “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, pp. 1021–1030.
- [30] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
- [31] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
- [32] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [33] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

- [34] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.
- [35] X. Chen, T. Jiang, J. Song, M. Rietmann, A. Geiger, M. J. Black, and O. Hilliges, “Fast-snarf: A fast deformer for articulated neural fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [36] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, “Neuman: Neural human radiance field from a single video,” in *Proceedings of the European conference on computer vision (ECCV)*, 2022.
- [37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.
- [38] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.
- [39] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.
- [40] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.
- [41] J. Gao, W. Chen, T. Xiang, A. Jacobson, M. McGuire, and S. Fidler, “Learning deformable tetrahedral meshes for 3d reconstruction,” *Advances In Neural Information Processing Systems*, vol. 33, pp. 9936–9947, 2020.
- [42] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, “Collaborative regression of expressive bodies using moderation,” in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 792–804.
- [43] A. Doi and A. Koide, “An efficient method of triangulating equi-valued surfaces by using tetrahedral cells,” *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 1, pp. 214–224, 1991.
- [44] Y. Xiu, J. Yang, X. Cao, D. Tzionas, and M. J. Black, “ECON: Explicit Clothed humans Optimized via Normal integration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

时间安排与分工统计表

组员信息（含组长）			
学生姓名	王威	学号	521021910283
项目分工	3D 模型构建与姿态识别部分的开发与实验		
学生姓名	周旭东	学号	521021910829
项目分工	三维人脸构建和二维换脸模型尝试、数据分析、方案确定		
学生姓名	代俊豪	学号	521021911093
项目分工	二维面部识别与人脸模型转换		
时间安排/ Schedule	2023年11月10日		
	↓ 初步选题与文献阅读 基础路线规划		
	2023年12月1日		
	↓ 多种模型尝试对比 最终方案制定		
	2023年12月24日		
	↓ 模型数据处理 模型创新优化 撰写总结报告		
	2024年1月6日		
	↓ 总结汇报 提交报告		
2024年1月7日			