目录

[1 绪论 3](#_Toc134541224)

[1.1 项目背景与意义 3](#_Toc134541225)

[1.2 研究与产业现状 3](#_Toc134541226)

[1.2.1 研究现状 3](#_Toc134541227)

[1.2.2 产业现状 4](#_Toc134541228)

[1.2.3 现状分析与项目目标 4](#_Toc134541229)

[1.3 文章结构 4](#_Toc134541230)

[2 模型搭建中的关键问题与解决方案 5](#_Toc134541231)

[2.1 关键问题分析 5](#_Toc134541232)

[2.1.1 深度学习模型搭建 5](#_Toc134541233)

[2.1.2 数据集选择 5](#_Toc134541234)

[2.1.3 实际场景展示 5](#_Toc134541235)

[2.2 方案设计 5](#_Toc134541236)

[2.3 技术选型 5](#_Toc134541237)

[3 手部运动估计模型设计 6](#_Toc134541238)

[3.1 环境说明与搭建 6](#_Toc134541239)

[3.1.1 环境选择 6](#_Toc134541240)

[3.1.2 环境配置 6](#_Toc134541241)

[3.2 Artiboost模型 6](#_Toc134541242)

[3.2.1 模型训练与测试 6](#_Toc134541243)

[3.2.2 输入部分 6](#_Toc134541244)

[3.2.3 输出部分 6](#_Toc134541245)

[3.3 TriHorn模型 6](#_Toc134541246)

[3.3.1 模型训练与测试 6](#_Toc134541247)

[3.3.2 输入部分 6](#_Toc134541248)

[3.3.3 输出部分 6](#_Toc134541249)

[3.3.4 模型优化 6](#_Toc134541250)

[4 应用软件设计 7](#_Toc134541251)

[4.1 模块设计 7](#_Toc134541252)

[4.1.1 数据集生成 7](#_Toc134541253)

[4.1.2 Socket通信与优化 7](#_Toc134541254)

[4.1.3 摄像画面截取 7](#_Toc134541255)

[5 结果展示与分析 7](#_Toc134541256)

[5.1 模型训练结果 7](#_Toc134541257)

[5.2 模型对比实验结果 7](#_Toc134541258)

[5.3 数据集对比实验结果 7](#_Toc134541259)

[6 总结 7](#_Toc134541260)

[6.1 总结 7](#_Toc134541261)

[6.2 趋势展望 7](#_Toc134541262)

# **1 绪论**

## 项目背景与意义

手势作为一种新型的交互手段可以实际应用于一些具有交互性质的场景，例如：文化遗址、博物馆等等。传统的展馆多数是以图片和文字作为交互的载体，能够传达给的信息十分有限。可能由于时代或环境因素，人们无法体会到图片和文字信息背后所包含的内容，而手势作为一种新型的交互工具，可以很大程度地弥补图片和文字的不足。随着技术的发展以及人们对于信息获取需求的不断多样化，基于AR、VR的交互式体验逐渐开始流行起来。

而交互最大的难点就在于姿态估计，所以如何提升手部信息的精准性能够大幅提高交互式体验在实际场景的可行性。在现有的体感交互设备中，大多都是专业且昂贵的动作捕捉设备，可以通过识别人体动作、手势等信息，实现对计算机的控制，从而实现更加自然、只管的人机交互方式，通常需要在手部绑定特定的传感器来识别手势。而神经网络依靠训练数据来不断学习和随时间推移提高自身精度。因而可以做到无需专业设备，通过训练得出的网络来估计手势姿态，大大降低了交互成本，同时还具有较高的泛用性。

虚拟手部生成数据集的方法能够快速、高效地获取到大量的手部数据，避免了复杂的专业设备的使用。这些虚拟空间手部数据同样具有有效的手部信息，能够快速地在虚拟空间中生成想要的数据集，避免了现实数据采集的限制和约束。

## 1.2 研究与产业现状

### **1.2.1 研究现状**

随着深度学习、神经网络和图像处理的不断发展，近年来，计算机视觉领域出现许多优秀的算法架构，使得计算机视觉技术正在广泛被使用，从自动驾驶汽车到医疗保健再到安全系统，计算机能够做一些更为精准的数据预测与数据增强。

**卷积神经网络。**卷积神经网络避免了对图像的复杂前期预处理，可以直接输入原始图像。Alex Krizhevsky、Hinton等人在2012年提出的AlexNet[1], 首次提出了Relu激活函数以及最大化池。相比于Sigmoid激活函数，训练效率成倍提升，从而获得了当时ImageNet[2]竞赛冠军。Kaiming He等人于2015年提出的Resnet[3], 解决了深度神经网络的“退化”问题，即使用浅层直接堆叠成深层网络，不仅难以利用深层网络强大的特征提取能力，而且准确率会下降。由于Resnet[3]的泛用性较高，因此许多数据增强模型会优先选择Resnet作为预训练模型，减少大量的训练资源和时间。

**对比学习**。对比学习是无监督学习中的重要分支。采用监督学习方法训练神经网络，需要采集大规模的数据集，但是数据集收集的成本非常高，例如用于深度卷积神经网络训练的ImageNet数据集中包含约130万张标签图像，人工标记这些图像需要花费大量的时间。 用于训练 卷积神经网络进行视频动作识别的Kinetics数据集中包含50万个视频，且每个视频持续时间约10s，这需要花费更长的时间对数据集进行注释。无监督学习孕育而生，解决了训练的数据集问题。在2019年MoCo[4]的提出使得原先不被看好的无监督对比学习一下子成为了最热门的研究方向之一。MoCo模型无需提前标注数据集的正确语义信息，计算机只需要区别数据之间的是否类似，从而开拓了对比学习模型。在之后的2020年，随着SimCLR[5]、SWaV[6]更多的框架被提出，人们发现无监督对比学习是可行且高效的，相比于原先的有监督学习需要大规模的数据集以及非常耗时的数据集标注来说，无监督学习就很好的避免了这些缺点，通过学习数据中的语义信息、特征信息来不断学习，从而应用在下游的任务中。

### 1.2.2 产业现状

由于计算机视觉现阶段主要还停留在科研阶段，实际实质化产出内容较少，因此存在一个巨大的市场缺口。实际场景中许多地方都需要使用计算机视觉相关的技术，例如：工业领域、医疗诊断、图像识别等等。在人脸识别领域，通过计算机进行支付验证、人脸跟踪的技术已经相对成熟。在自然语言处理领域，由于gpt-3[7]的出现，基于深度学习的大模型基本已经能够与人正常交流，并开始逐步进入了市场应用，被人们广泛使用。现今随着类似于gpt-3[7]的大模型越来越流行，未来会有更多领域使用大模型的方法进行训练和测试，其中也包括计算机视觉。但是由于手势的复杂性导致目前暂时没有一个非常优秀的网络模型来解决姿态估计的问题，因此整个手势识别领域现阶段还只是处于研究阶段。

### 1.2.3 现状分析与项目目标

近年来，在学术领域提出的许多前瞻性的新方法，造就了计算机视觉的飞速发展，但是由于多数模型的研究都只停留在理论层面的结论，没有考虑在实际场景中的应用，从而导致缺少与工程领域的对接，需要做大量的修改才能满足工程上的需求。因此我们的方法是在前人的基础上进行重构，使用其模型的核心思想：整个神经网络模型，然后将整个模型进行封装，从而能够应用到实际场景中。

本项目的目标是构造一个面向实际场景的应用。首先需要建立数据集生成模型，该模型使用骨骼动画化驱动虚拟手渲染生成数据集并且针对神经网络模型的输入要求批量采集数据。基于如上数据集，交由手部-物体姿态估计算法模型进行训练，然后封装成一个手部姿态估计工具。最后应用于下游的实际应用中，根据输入的手势数据在应用的虚拟场景中实时显示和交互。

## 1.3 文章结构

本文总共分为六个章节

第一章：绪论。

第二章：关键问题分析与解决

第三章：神经网络模型设计

第四章：模块设计

第五章：实验对比与分析

第六章：总结与展望

# 2 模型搭建中的关键问题与解决方案

## 2.1 关键问题分析

### 2.1.1 深度学习模型搭建

### 2.1.2 数据集选择

### 2.1.3 实际场景展示

## 2.2 方案设计

## 2.3 技术选型

# 3 手部运动估计模型设计

## 3.1 环境说明与搭建

### 3.1.1 环境选择

### 3.1.2 环境配置

## 3.2 Artiboost模型

### 3.2.1 模型训练与测试

### 3.2.2 输入部分

### 3.2.3 输出部分

## 3.3 TriHorn模型

### 3.3.1 模型训练与测试

### 3.3.2 输入部分

### 3.3.3 输出部分

### 3.3.4 模型优化

# 4 应用软件设计

## 4.1 模块设计

### 4.1.1 数据集生成

### 4.1.2 Socket通信与优化

### 4.1.3 摄像画面截取

# 5 结果展示与分析

## 5.1 模型训练结果

## 5.2 模型对比实验结果

## 5.3 数据集对比实验结果

# 6 总结

## 6.1 总结

## 6.2 趋势展望

参考文献

1. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
2. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
3. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
4. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.
5. Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.
6. Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments[J]. Advances in Neural Information Processing Systems, 2020, 33: 9912-9924.
7. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.