



南方科技大学  
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

# 《电子科学创新实验 II》 课程报告

题    目： 基于视觉方法的手语翻译系统  
姓    名： 应逸雯  
学    号： 12210159  
系    别： 电子与电气工程系  
专    业： 信息工程  
指导教师： 张宏

2024 年 06 月 15 日



# 基于视觉方法的手语翻译系统

应逸雯

(电子与电气工程系 指导教师：张宏)

**[摘要]**：本项目完成了基于视觉方法的手语翻译系统的设计与实现，能够辅助聋哑人与健听人的交流。该系统采集 RGB 图像，利用图像预处理算法和 CNN、ViT 深度学习技术，实现对汉语手语的静态手势的自动识别。手势识别系统支持 30 个汉语拼音手势、10 个汉语数字手势，在无遮挡情况中，汉语拼音识别准确率约 75%，数字识别准确率 100%，在有遮挡情况中，基本能够识别，表现出一定的鲁棒性和可靠性。识别结果经过拼接分割后转换为文本并以语音形式输出，形成完整的翻译系统。

**[关键词]**：手势识别；手语识别系统；图像处理；图像识别；RGB 图像；深度学习。

# 目录

1. 项目背景.....	1
2. 项目研究内容.....	2
2.1 图像预处理算法.....	2
2.1.1 手部前景提取.....	2
2.1.2 裁切感兴趣区域.....	3
2.2 深度学习模型.....	4
2.2.1 卷积神经网络.....	4
2.2.2 视觉自注意力模型.....	4
2.3 图像数据集增强.....	5
3. 项目实施过程.....	6
3.1 训练数据集准备.....	6
3.2 图像预处理.....	8
3.3 深度学习模型预测图像类别.....	9
3.4 文本拼接并以语音形式输出.....	10
4. 项目方案.....	10
4.1 传统方法手势分类：二维约束法计算向量夹角.....	10
4.2 二维卷积神经网络深度学习模型.....	11
4.2.1 以图像灰度值作为单通道输入数据.....	11
4.2.2 以图像 HSV 值作为三通道输入数据.....	12
4.3 一维卷积神经网络深度学习模型-手部骨架坐标信息.....	14

4.3.1 数字手势分类问题.....	14
4.3.2 拼音手势分类问题.....	15
4.4 Vision-Transformer 深度学习模型.....	15
4.4.1 数字手势分类问题.....	16
4.4.2 拼音手势分类问题.....	16
4.5 三维神经网络.....	17
4.5.1 从 RGB 图像估计三维坐标的应用.....	18
5. 项目总结.....	18
参考文献.....	19

## 1. 项目背景

手语作为聋哑人的沟通方式，保障了聋哑人的生活得以正常进行。然而，健听人中绝大部分人不掌握手语，无法理解聋哑人的信息。因此，一套完善的手语翻译系统能够帮助解决这一问题。手语识别也可拓展至其他的手势指令相关任务中，如智能家居、人机交互、虚拟现实等领域。手语翻译系统主要任务包括手语识别和文本翻译。本项目主要聚焦于手语识别部分，文本翻译主要通过调包和参照他人方法完成以形成完整的系统。

目前，手语识别主要采用传感器手套和视觉方法两种方式。传感器手套方式由于获取信息丰富、准确率高、识别速度快而备受青睐。然而，由于其价格昂贵且存在操作约束性的问题，逐渐被视觉方法所取代。

视觉方法主要分为静态手势识别和动态手势识别两类。静态手势识别针对单帧图像进行处理，主要解决准确率、光照和角度变化等问题。动态手势识别则涉及连续帧的时序关系和孤立词之间的过渡帧问题。当前的手势识别方法主要包括传统方法和基于深度学习的方法。传统方法受制于设计的算法，泛化性较差，逐渐被深度学习方法所取代。深度学习方法中，静态手势识别主要采用卷积神经网络，而动态手势识别则利用 3D 卷积神经网络、循环神经网络和长短时记忆网络等技术[1]。近年来 Vision Transformer 也逐渐被应用于图像分类等计算机视觉任务中[2]。其将图像视为序列数据，并利用 Transformer

的注意力机制来建模图像中的全局和局部关系。

静态手势识别的图像来源主要包括 RGB 相机、深度相机、红外相机和 RGBD 相机等。RGB 相机提供彩色图像，具有较高的分辨率和广泛的应用，但对深度信息的获取有限。深度相机能够获取场景的深度信息，但在复杂场景下易受到光照和材质影响。红外相机在低光条件下表现良好，但对于彩色信息的获取较为有限。RGBD 相机结合了 RGB 相机和深度相机的优点，提供了丰富的彩色和深度信息，但成本相对较高。

目前，使用 RGBD 相机识别手势已能达到约 90% 及以上的准确率 [3]。而相较而言，获取较少信息的 RGB 相机实惠易得，但由于丢失了深度信息，识别准确率普遍不高，如能提高使用 RGB 相机场景的准确率，有利于手势识别算法应用于现实场景中。近年来，通过 RGB 信息估计深度信息以弥补 RGB 相机的不足也逐渐应用于手势识别中 [4]。

鉴于系统的可扩展性、实用性和可行性等因素，本研究选择采用 RGB 相机获取图像数据，并基于深度学习模型解决汉语手语的静态手势识别问题。

## 2. 项目研究内容

### 2.1 图像预处理算法

#### 2.1.1 手部前景提取

使用 YCrCb 颜色空间判定肤色，即通过明暗信息、红色色度分量、蓝色色度分量信息分离肤色区域。前景图像皮肤在 YCrCb 颜色空间中的 Cr 分量上有显著特征，区分效果优于 RGB 颜色空间。

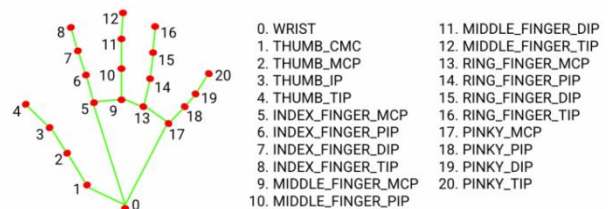
使用 OTSU 算法确定阈值，通过 Cr 分量提取手部。OTSU 算法是一种自适应的阈值分割算法，通过最大化类间方差来确定最优阈值，将图像划分为前景和背景两部分。类间方差定义为  $\sigma_b^2(T) = \omega_0(T) \cdot \omega_1(T) \cdot (\mu_0(T) - \mu_1(T))^2$ ，其中  $\omega_0, \omega_1$  为由灰度概率定义的权重， $\mu_0, \mu_1$  为背景和前景的均值。

通过高斯模糊处理和形态学闭运算操作去除噪声和填补孔洞。保留三个最大的轮廓的区域以防止背景相似颜色干扰。使分割图像更大程度保留前景和去除背景。

### 2.1.2 裁切感兴趣区域

通过精准定位手部的关键点，可以确定手部的边界并裁切出感兴趣区域。这种方法不仅减少了背景干扰，还最大化了数据利用效率，加快了运行和训练速度。[5]

Google 的 Mediapipe 库提供的 API 能够良好地估计手部骨架位置，



尽管每个点可能存在误差，但在确定手部位置范围方面具有足够的精度。通过遍历如图的 21 个关键点，可以确定手部的位置范围，并向四周扩展 20 个像素，确保手部信息完全涵盖在裁切的感兴趣区域中。



## 2.2 深度学习模型

### 2.2.1 卷积神经网络

卷积神经网络（CNN）是一种常用于处理图像数据的深度学习模型。通过多层卷积和池化操作来提取图像的特征，通过全连接层进行分类或回归。

在卷积层中，通过将一个卷积核应用于输入图像的每个位置来提取特征。卷积操作可以表示为：

$(I * K)(x, y) = \sum_{i=-k}^k \sum_{j=-k}^k I(x+i, y+j) \cdot K(i, j)$ 。其中， $I$  表示输入图像， $K$  表示卷积核， $(x, y)$  表示输出图像中的位置， $(i, j)$  表示卷积核的索引， $k$  是卷积核的大小。

池化层用于减少特征图的尺寸，降低计算量，并保留最重要的信息。常用的池化操作包括最大池化和平均池化。

通过多次堆叠卷积和池化层，网络可以学习到更复杂的特征，最后通过全连接层进行分类或回归，得到训练模型。

### 2.2.2 视觉自注意力模型

Vision Transformer 是一种基于 Transformer 架构的图像处理模型，使用自注意力机制来处理图像数据。ViT 将图像划分为一系列固定大小的图块，并将这些图块视为序列输入到 Transformer 模型中

以训练模型。

首先，将输入图像分割成大小为  $N \times N$  的图块，并将每个图块展平为一维向量。然后，这些图块向量通过线性投影映射到更高维的嵌入空间中。即， $z_0 = [x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos}$ ，其中  $x_p^i$  表示第  $i$  个图块， $E$  为投影矩阵， $E_{pos}$  为位置编码。

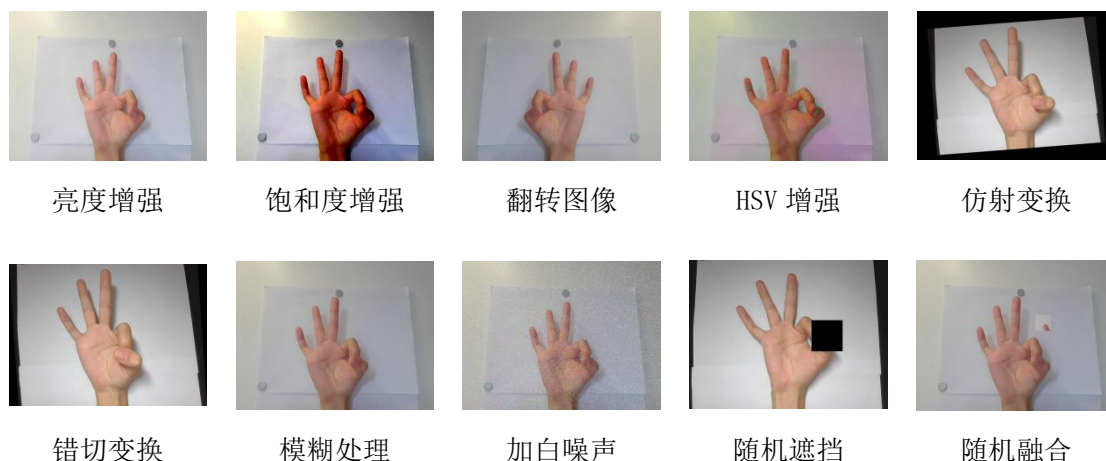
嵌入序列输入到标准的 Transformer 编码器中，通过多头自注意力机制和前馈神经网络进行处理。

经过多层 Transformer 编码器的处理后，输出特征被用于图像分类任务，通常通过全连接层进行分类。

## 2.3 图像数据集增强

通过对像素值的线性变换和非线性变换处理，可以对图像数据集增强，得到更多样的和鲁棒性更强的数据集。亮度增强和对比度增强操作采用像素值随机缩放方法，调整图像的亮度和对比度。图像旋转、翻转、反射、放大缩小图像和平移变换使用了仿射变换原理，在一定范围内随机构造旋转矩阵、翻转像素，与图像相乘，实现仿射变换。错切变换则是通过线性变换使图像中的每个点沿特定方向移动一定距离。HSV 数据增强随机调整了图像的色调、饱和度和亮度通道，改变了图像的外观。高斯模糊操作采用了卷积运算，利用高斯核对图像进行模糊处理。增加噪声通过生成服从特定分布的随机数并添加到图像中的像素值实现。随机遮挡通过使图像中一部分区域的像素值置为

(0, 0, 0) 来实现。随机融合通过随机选取另一张图片的一部分区域加入到当前图片的一部分区域模拟识别时的外界干扰[6]。组合迭代算法使数据集更多样，覆盖录制数据集所不能得到的各种极端场景。

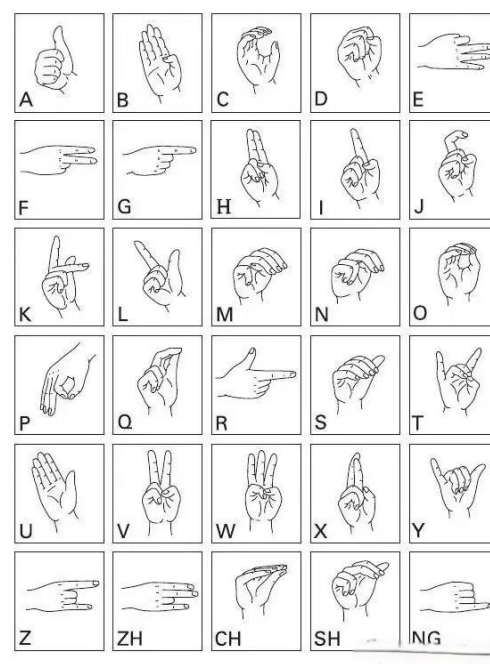


### 3. 项目实施过程

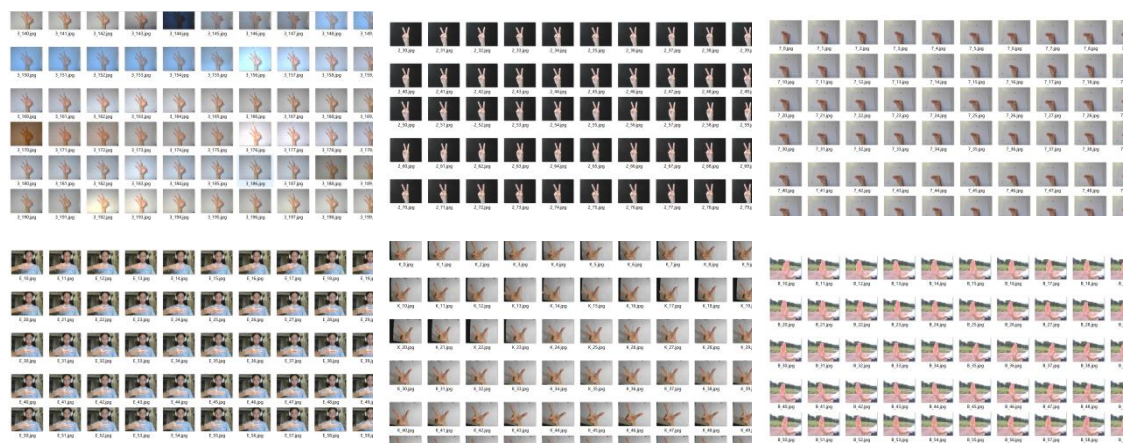
#### 3.1 训练数据集准备

由于网络资源中中国手语数据不丰富、多为 RGBD 数据和视频流等原因，不符合本项目所研究的基于 RGB 静态图像的汉语手语识别问题，故采用自建数据集进行训练和测试。

自建的数据集依照汉语手语规范动作制作。

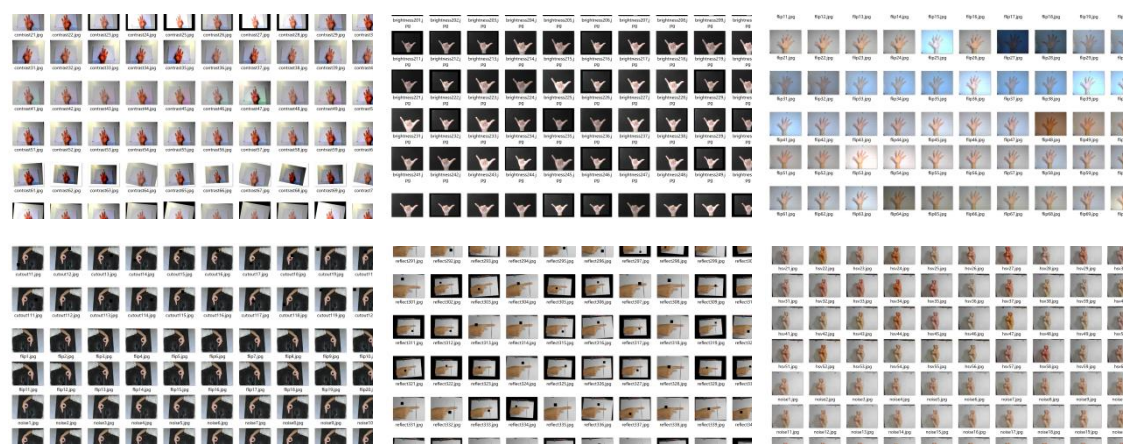


录制了千余张不同背景下（白色干净背景、黑色干净背景、室内杂乱背景、户外背景）、不同光照下（温和日光灯、室外强光、不同颜色补光）、不同角度下（录制时的手部转动）的手势。



录制结果示例


为应对更多复杂环境的情况，采用图像数据集增强算法扩充数据集。对图像进行水平翻转（模拟左右手）、对比度增强、亮度增强、仿射变换、错切变换、HSV 增强、模糊化、增加白噪声、增加随机遮挡、增加随机图像融合。增加数据集多样性和系统鲁棒性。



图像数据集增强结果示例

### 3.2 图像预处理

基于 YCrCb 颜色空间，使用 OTSU 算法区分前景和背景，提取手部信息，去除不必要的背景信息，减少背景干扰。使用 Mediapipe 库的 API 接口，得到手部骨骼点的位置信息，定位手部位置范围，裁切出手部区域。通过以上操作，处理的数据基本只包含了如图



的手部区域。对图像进行缩放，调整成 56\*56 像素的正方形大小，保证后续处理不同类型数据时不会出错。

对于 CNN 网络，无法直接输入图像数据，需要将信息按一定逻辑处理成数据。

可以保留每个像素的灰度值信息，在网络中进行一通道输入，计算速度快，但丢失了大量有用信息，也因而准确率不高。

为更大程度保存图像信息，可以保留每个像素的 HSV 值，在网络中进行三通道输入，计算速度慢，但准确率高。HSV 值能够基本恢复出图像，失真少，只存在不影响计算效果的图像拉伸，效果如图。

在简单手势中，也可以记录骨骼点坐标信息，保存 21 个骨骼点的 xy 坐标。计算速度非常快，在简单手势中表现很好，但鲁棒性低，存在干扰时骨骼点识别易出错，复杂手势、相似手势骨骼点相近。



### 3.3 深度学习模型预测图像类别

CNN 模型：根据每种任务，定义三层卷积层，卷积核大小分别为  $(7, 7)$ ， $(5, 5)$ ， $(3, 3)$ ，输出通道数分别为 35, 20, 10。激活函数使用了 LeakyReLU。训练一定数量的 epoch（保存了每一次训练结果，取最优的权重文件用于预测）。如图是训练结果。

```
C:\Windows\System32\cmd.e  X + v
164/164 [=====] - 90s 548ms/step - loss: 0.1633 - accuracy: 0.9566 - val_loss: 1.6391 - val_acc
uracy: 0.3898 - lr: 0.0010
Epoch 2/6
164/164 [=====] - 98s 595ms/step - loss: 0.0032 - accuracy: 0.9997 - val_loss: 0.6777 - val_acc
uracy: 0.8676 - lr: 0.0010
Epoch 3/6
164/164 [=====] - 111s 676ms/step - loss: 0.0013 - accuracy: 1.0000 - val_loss: 0.0757 - val_acc
uracy: 0.9962 - lr: 0.0010
Epoch 4/6
164/164 [=====] - 115s 702ms/step - loss: 5.9634e-04 - accuracy: 1.0000 - val_loss: 0.0167 - va
l_accuracy: 0.9972 - lr: 0.0010
Epoch 5/6
164/164 [=====] - 114s 694ms/step - loss: 4.4512e-04 - accuracy: 1.0000 - val_loss: 0.0122 - va
l_accuracy: 0.9972 - lr: 0.0010
Epoch 6/6
164/164 [=====] - ETA: 0s - loss: 5.3236e-04 - accuracy: 0.9999
Epoch 00006: ReduceLROnPlateau reducing learning rate to 0.0005000000237487257.
164/164 [=====] - 114s 696ms/step - loss: 5.3236e-04 - accuracy: 0.9999 - val_loss: 0.0126 - va
l_accuracy: 0.9966 - lr: 0.0010
D:\wendy\study\2023-2024Spring\EE318\sign-language-detection-cnn-deeplearning>
```

ViT 模型：使用 vit\_base\_patch16\_224 模型，train 集：  
validation 集：test 集分割为 8:1:1。初始学习率设定为 0.001，学  
习率衰减因子设定为 0.01。冻结了底层权重，加快训练。训练一定  
数量的 epoch（保存了每一次训练结果，取最优的权重文件用于预测）。  
如图是训练结果。

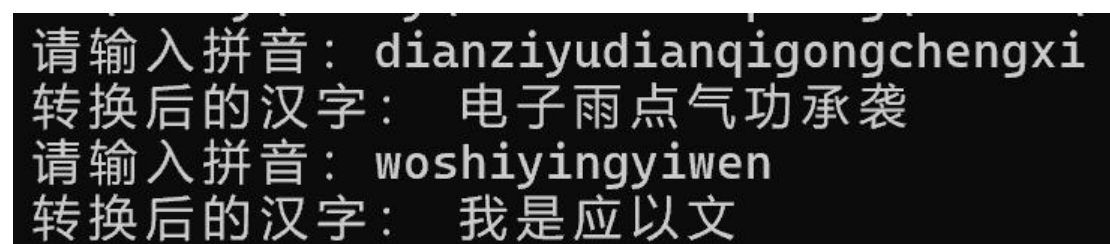
```
train epoch 0] loss: 0.853, acc: 0.814: 100% | 6774/6774 [09:37<00:00, 11.74it
valid epoch 0] loss: 0.089, acc: 0.971: 100% | 847/847 [01:10<00:00, 11.98it/s
train epoch 1] loss: 0.453, acc: 0.901: 100% | 6774/6774 [09:31<00:00, 11.85it
valid epoch 1] loss: 0.166, acc: 0.959: 100% | 847/847 [01:10<00:00, 11.96it/s
train epoch 2] loss: 0.315, acc: 0.927: 100% | 6774/6774 [09:26<00:00, 11.97it
valid epoch 2] loss: 0.031, acc: 0.989: 100% | 847/847 [01:10<00:00, 11.97it/s
train epoch 3] loss: 0.217, acc: 0.947: 100% | 6774/6774 [09:24<00:00, 11.99it
valid epoch 3] loss: 0.037, acc: 0.986: 100% | 847/847 [01:10<00:00, 11.98it/s
train epoch 4] loss: 0.170, acc: 0.956: 100% | 6774/6774 [09:25<00:00, 11.98it
valid epoch 4] loss: 0.016, acc: 0.995: 100% | 847/847 [01:10<00:00, 12.03it/s
valid epoch 4] loss: 0.014, acc: 0.996: 100% | 847/847 [01:10<00:00, 11.97it/s
est accuracy: 0.9959, Test loss: 0.0135
```

### 3.4 文本拼接并以语音形式输出

图像识别结果在经过验证后加入缓冲队列，以防止错误识别影响结果输出。验证过程包括以下步骤：确保至少连续两次相同的结果、确保与上一个存入缓冲队列的字符不同、并且仅记录连续识别到的结果一次。

依据汉语声母韵母音节规律，对拼音做简单分割。分割后的拼音按照列表形式返回。当识别到下一个字时，输出上一个字，以保证分割的完整性。

使用 Pinyin2Hanzi 库的 API 将拼音转换为汉字。尽管转换后的汉字可能与实际想表达的意思有所出入，但在可接受范围内，不影响语音输出效果。使用 pyttsx3 库的 API 对生成的汉字进行语音输出。为防止语音输出阻塞图像识别任务，增加了一个专门的语音任务线程进行语音播放。如图该步骤拼音转汉字和语音的效果。



```
请输入拼音： dianziyudianqigongchengxi
转换后的汉字： 电子雨点气功承袭
请输入拼音： woshiyingyiwen
转换后的汉字： 我是应以文
```

拼音转汉字和语音效果示例

## 4. 项目结果

### 4.1 传统方法手势分类：二维约束法计算向量夹角

使用 Google 的 Mediapipe 库 API 提供的手部坐标模型，计算出手指指尖与手掌夹角（以拇指为例， $(0, 1)$  向量和  $(3, 4)$  向量的夹角）。通过二维约束法判定手势类型。

由于角度与试验者手型高度相关，且使用的是手工设计的计算方法，利用夹角是否落在范围内，判断手指开合。在简单手势时能得到正确结果，但鲁棒性非常低，且无法拓展至复杂手势（如数字 9 无法判断食指开合，与 0 情形会相同）。目前也存在各类手工设计算法，但整体上不敌由大量数据驱动的深度学习模型的准确率和泛化性。

在测试中，识别 0, 1, 2, 3, 4, 5, 6, 8（该算法不支持数字 7 和 9 的识别），总计准确率为 88.57%。各标签准确率如下表：

标签	0	1	2	3	4	5	6	8
正确数量	3325	3412	3481	1612	3575	3583	3490	3032
样本总量	3600	3600	3600	3600	3600	3600	3600	3600
正确率	0.9236	0.9478	0.9669	0.4478	0.9931	0.9953	0.9695	0.8422

## 4.2 二维卷积神经网络深度学习模型

### 4.2.1 以图像灰度值作为单通道输入数据

对图像预处理后，使用每个像素点的灰度值信息进行训练和预测。

使用了 CNN 模型，使用了三层卷积层，卷积核大小分别为  $(7, 7)$ ， $(5, 5)$ ， $(3, 3)$ ，输出通道数分别为 35, 20, 10。激活函数使用了



LeakyReLU。训练了 6 个 epoch。

从训练结果尝试中发现，数字 0-9 已都可以识别，且能够一定程度的抗环境干扰，但准确率并非每次都能很高，偶有出错情况。出错可能由于手势相近，在仅有 56\*56 像素的图像上，叠加了背景干扰因素转成灰度值后，部分图像容易相近。而模型可能过拟合，故相近的手势容易被估计为其他情况。响应速度已可以接受。采用他人手势试验，绝大部分能得到正确结果，部分手势仍存在问题。

在数字识别任务中，总准确率为 72.03%。各标签准确率如下表：

标签	0	1	2	3	4	5	6	7	8	9
正确数量	293	389	347	300	364	326	380	11	333	137
样本总量	400	400	400	400	400	400	400	399	400	399
正确率	0.733	0.973	0.878	0.750	0.910	0.815	0.950	0.028	0.833	0.343

部分手势无法识别，考虑灰度值信息丢失了丰富的图像信息，考虑能够还原图像信息的其他方式。

#### 4.2.2 以图像 HSV 值作为三通道输入数据

以图像 HSV（色度、饱和度、亮度）数据作为输入数据，增大了计算量，但更大程度的保留了信息，以提高准确率。仍然以同样的 2D-CNN 模型进行训练。训练结果经尝试能够识别一些复杂手势。

在数字识别任务中，总准确率为 83.64%。各标签准确率如下表：

标签	0	1	2	3	4	5	6	7	8	9
正确数量	359	332	362	306	366	325	361	266	318	349
样本总量	400	400	400	400	400	400	400	399	400	399
正确率	0.898	0.830	0.905	0.765	0.915	0.813	0.903	0.667	0.795	0.875

在灰度信息作为输入数据的模型所不能良好识别的数字7和9中，识别效果已有大幅提升。

考虑拓展至类别繁多、手势复杂且相近的汉语手语识别问题。

在汉语手语静态手势三十分类问题中，总准确率为 52.0%。各标签准确率如下表：

标签	A	B	C	CH	D	E	F	G	H	I
正确数量	0	7	125	126	76	64	107	93	20	130
样本总量	135	137	138	137	138	126	128	127	136	140
正确率	0.000	0.051	0.906	0.920	0.551	0.508	0.836	0.732	0.147	0.928
标签	J	K	L	M	N	NG	O	P	Q	R
正确数量	39	130	124	58	8	56	5	114	1	127
样本总量	140	140	140	140	140	136	140	140	139	138
正确率	0.279	0.929	0.886	0.414	0.057	0.412	0.036	0.814	0.007	0.920
标签	S	SH	T	U	V	W	X	Y	Z	ZH
正确数量	5	13	110	76	68	136	129	137	38	3
样本总量	134	138	135	140	140	140	140	140	126	123
正确率	0.037	0.094	0.815	0.543	0.486	0.971	0.921	0.979	0.302	0.024

观察数据发现，许多手势几乎无法识别。其原因可能主要由于多类别之间手势非常相似（如 B 与 U）；手部遮挡关系对于人眼判断其类别极为重要，但对于仅能从一个角度观察到 RGB 图像的识别系统，无法判定关系，进而容易分类出错。

#### 4.3 一维卷积神经网络深度学习模型-手部骨架坐标信息

##### 4.3.1 数字手势分类问题

对于能够清晰定义骨架信息的情况，使用骨架坐标信息能够更好地模拟人眼判断手势的逻辑。前人研究经验中同样发现，由骨骼信息判定手势相较于像素值而言，能够有更好的效果[5]。

使用 Google 的 Mediapipe 库所支持的手部坐标模型 API，定位手部 21 个关键点的位置，以关键点二维坐标信息输入至模型中进行训练和预测。

由于所需的预处理后的数据较像素值比较法显著减少，计算量小，使用骨骼坐标的方法在训练速度、响应速度上也有显著优势。

在良好的测试集（不刻意遮挡、温和光照环境）下，该方法在数字识别的所有标签（0, 1, 2, 3, 4, 5, 6, 7, 8, 9）中达到了 100%的准确率。可能由于个人所能测试的数据集有限，无法穷尽所有情况，但可以证明该方法在简明手势中能够有优秀表现。

### 4.3.2 拼音手势分类问题

依据骨架坐标信息的处理算法在数字识别中的优秀表现，尝试扩展该模型至汉语手语识别中。测试总准确率为 78%。以下是各标签的具体准确率：

标签	A	B	C	CH	D	E	F	G	H	I
正确率	0.98	1.00	0.99	1.00	0.46	0.84	0.99	1.00	0.18	0.80
标签	J	K	L	M	N	NG	O	P	Q	R
正确率	0.54	0.61	0.84	0.97	0.99	1.00	0.21	0.91	0.31	0.98
标签	S	SH	T	U	V	W	X	Y	Z	ZH
正确率	0.00	0.97	0.52	0.90	0.97	1.00	0.96	0.99	0.92	0.85

对于汉语拼音的三十个手势，整体得到了良好的准确率，虽然相当一部分手势准确率未达到很高（ $<0.95$ ），但整体具有可靠性。

需要注意的是，使用骨架坐标的方法对于预测数据的规范性要求强，若出现极强/极弱光照、遮挡等情况，无法正确识别到手部骨架位置，该方法容易迅速失效。

## 4.4 Vision-Transformer 深度学习模型

Transformer 是一种用于序列建模的深度学习模型架构，在自然语言处理中实现了巨大突破。其自注意力机制能够有效处理所有位置的信息，捕捉长距离依赖关系。将视觉任务应用于 Transformer 模型

也有优秀的效果，具有更好的全局信息捕捉能力、可解释性、适应性。在图像分类中，Vision Transformer 模型已被证实可以得到与 CNN 相当甚至超过 CNN 的性能[6]。

但 ViT 也有一定缺点，其运算量较 CNN 而言较大，训练速度和响应速度较慢，不过可以通过冻结权重加快训练速度。ViT 需要庞大的数据集支撑其运算效果，对于小数据集效果不佳。

#### 4.4.1 数字手势分类问题

将 Vision Transformer 用于数字手势的识别中，加载了预训练权重，使用 vit\_base\_patch16\_224 模型。取第四个 epoch 权重文件效果最佳。测试得到总准确率为 89.24%。下表为各标签对应准确率：

标签	0	1	2	3	4	5	6	7	8	9
正确率	0.447	1.00	0.72	1.00	0.99	1.00	1.00	1.00	1.00	0.83
正确样本	42	100	72	100	99	100	100	100	100	83
样本总量	94	100	100	100	100	100	100	100	100	100

可能由于训练样本数量较少，表现结果不佳，但在可接受范围内。且 Vision Transformer 最适用于类别更多的任务，可能不适合这一情形。

#### 4.4.2 拼音手势分类问题

使用同样的模型对汉语拼音识别问题实践。取第四个 epoch 权重

文件效果最佳。测试总准确率为 75%。下表是各标签的准确率：

标签	A	B	C	CH	D	E	F	G	H	I
正确率	0.436	0.000	1.000	0.993	1.000	0.942	0.957	0.949	0.863	0.236
正确样本	61	0	140	137	140	130	132	131	120	33
样本总量	140	139	140	138	140	138	138	138	139	140
标签	J	K	L	M	N	NG	O	P	Q	R
正确率	0.771	0.614	0.557	0.243	0.086	0.950	0.986	1.000	1.000	0.964
正确样本	108	86	78	34	12	133	138	140	140	135
样本总量	140	140	140	140	140	140	140	140	140	140
标签	S	SH	T	U	V	W	X	Y	Z	ZH
正确率	0.906	0.243	0.957	1.000	0.550	1.000	1.000	0.943	0.986	0.544
正确样本	125	34	134	140	77	140	140	132	136	75
样本总量	138	140	140	140	140	140	140	140	138	138

在部分手势中，仍然基本识别失败（B, I, M, N），考虑仍是由于手势过于相近的问题。其余大部分类别中，基本达到了 95%+的准确率，有良好的预测效果。

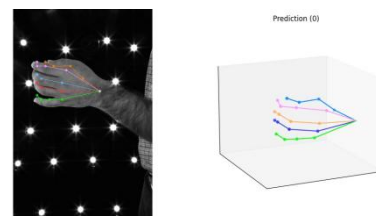
#### 4.5 三维神经网络

RGB 图像相较于 RGB-D 图像而言，缺少了深度信息，而深度信息作为人眼判断手势类别的重要依据，影响了手势识别的准确率。通过从 RGB 图像恢复深度信息，可以弥补这一问题。目前已有一定数量的

研究旨在从 RGB 图像恢复深度信息[7]。

#### 4.5.1 从 RGB 图像估计三维坐标的应用

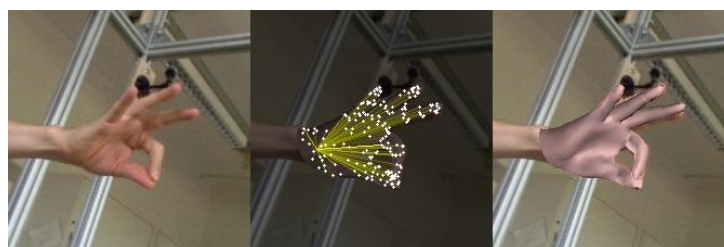
使用了 open-mmlab 的 mmpose 库 API 接口，效果如图，能够从 RGB 图像恢复出手部骨骼的三维坐标信息（XYZ）。



将如图点的三维坐标信息输入至 CNN 网络中，应用于汉语拼音的手势识别问题，训练数据，测试得到准确率为 62.70%。

考虑参照如图的手部点连接关系，使用图卷积神经网络，丰富输入信息，增强准确率。（未实现）

在测试中发现部分手势不能够准确恢复出手部骨骼点 XYZ 坐标信



息，导致准确率较低。考虑涵盖全手的深度信息的架构[8]，效果如图，结合 RGB 信息一同输入深度学习模型中，最大化数据使用效率，提升识别准确率。（未实现）

## 5. 项目总结

对于简单手势（数字手势），骨骼坐标信息点输入一维卷积神经网络得到了非常好的效果。对于复杂手势，图像输入 ViT 模型得到了可观的效果。对于遮挡问题，在训练图像中使用 CUT-OFF, CUT-MIX 图

像数据集增强,能够有效提高鲁棒性。对于不同光照、不同角度的扩展,在训练数据集中增加相应的操作录制。对于背景干扰,通过提取前景和 ROI 分割算法,使预测聚焦于手部信息,滤除了背景。

与前人研究相比,本项目实现了采用 RGB 相机获取图像的汉语手语识别任务;实现了 ViT 模型在汉语手语识别中的应用;实现了从手势转换至语音输出的完整系统;实现了三维姿态估计在手势识别中的应用。

项目搭建了手语翻译系统,仅采用电脑手机都能支持的普通 RGB 相机,使用 ViT 网络识别汉语手语拼音,并于队列中处理输出语音,实现了从手势到语音的完整系统。准确率、鲁棒性可靠,实用性强。

项目未来还可以拓展至使用 3D 姿态估计,从 RGB 图像还原深度信息,使用更丰富的信息增强系统的准确性。项目还可以将成果封装成 APP,实现该系统的实用功能。

## 参考文献

- [1] 张淑军,张群,李辉. 基于深度学习的手语识别综述[J]. 电子与信息学报, 2020, 42(04):1021-1032.
- [2] Zhang Y, Wang J, Wang X, et al. Static hand gesture recognition method based on the vision transformer[J]. Multimedia Tools and Applications, 2023, 82(20): 31309-31328.
- [3] TANG Ao, LU Ke, WANG Yufei, et al. A real-time hand posture recognition system using deep neural networks[J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(2): 1-23.



- [4] 李琦. 基于 GCN 的 RGB 三维手势跟踪算法研究[D]. 内蒙古科技大学, 2023.
- [5] Devineau G, Moutarde F, Xi W, et al. Deep learning for hand gesture recognition on skeletal data[C]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 106-113.
- [6] Zhang Y, Wang J, Wang X, et al. Static hand gesture recognition method based on the vision transformer[J]. Multimedia Tools and Applications, 2023, 82(20): 31309-31328.
- [7] Zimmermann C, Brox T. Learning to estimate 3d hand pose from single rgb images[C]. Proceedings of the IEEE international conference on computer vision. 2017: 4903-4911.
- [8] Kanazawa A, Black M J, Jacobs D W, et al. End-to-end recovery of human shape and pose[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7122-7131.