Journal of Image and Signal Processing 图像与信号处理, 2024, 13(3), 338-347
Published Online July 2024 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2024.133029

Hans汉斯

基于改进3D-ResNet34模型的视频人员情绪识别研究

郭复澳，姚克明*，王中洲

江苏理工学院电气信息工程学院，江苏 常州

收稿日期：2024年6月17日；录用日期：2024年7月7日；发布日期：2024年7月18日

摘 要

由于人员情绪的多变性，面部表情会随着时间的推移与情绪的变化而变化，这一点在视频数据中表现尤为显著，此时使用只对单帧图像进行训练的网络模型来判别视频数据中的人员心情显然不可行。因此为解决上述问题，提出一种结合3D-ResNet34、ConvLSTM以及Transformer的网络模型用于人员情绪识别的方法。首先，对视频数据集进行预处理，将完整的视频数据划分成若干个连续片段，使用3D-ResNet34网络对视频片段进行空间特征提取。其次，在网络模型中设计添加ConvLSTM模块用于从空间特征中提取更深层次的时间维度特征。然后，通过使用自设计的Transformer模块对每个视频片段生成一个注意力分布，用于加权融合各帧特征，得到一个表示整个视频片段的注意力特征。最后对提取完成的时间维度特征与注意力特征进行特征融合，形成一个新的特征表示，并送入Softmax层进行分类识别。实验表明，本文设计的方法取得了较好的识别准确率。

关键词

人员情绪识别，特征提取，Transformer，特征融合

Research on Video Personnel Emotion Recognition Based on Improved 3D-ResNet34 Model

Fu'ao Guo, Keming Yao*, Zhongzhou Wang

School of Electrical and Information Engineering Jiangsu University of Technology, Changzhou Jiangsu

Received: Jun. 17th, 2024; accepted: Jul. 7th, 2024; published: Jul. 18th, 2024

*通讯作者。

文章引用: 郭复澳, 姚克明, 王中洲. 基于改进3D-ResNet34模型的视频人员情绪识别研究[J]. 图像与信号处理, 2024, 13(3): 338-347. DOI: 10.12677/jisp.2024.133029

Abstract

Due to the variability of people's emotions, facial expressions will change with the passage of time and the changes of emotions, which is particularly significant in video data. At this time, it is obviously not feasible to use the network model trained only on a single frame image to identify people's moods in video data. Therefore, in order to solve the above problems, a method combining 3D-ResNet34, ConvLSTM and Transformer network model is proposed for human emotion recognition. Firstly, the video data set is preprocessed, the complete video data is divided into several continuous fragments, and the spatial features of the video fragments are extracted using 3D-ResNet34 network. Secondly, ConvLSTM modules are added to the network model to extract deeper temporal features from spatial features. Then, a self-designed Transformer module can be used to generate an attention distribution for each video clip, which can be used to weight and fuse each frame feature to obtain an attention feature representing the entire video clip. Finally, the extracted time dimension features and attention features are fused to form a new feature representation, which is sent to the Softmax layer for classification and recognition. Experiments show that the method designed in this paper achieves a good recognition accuracy.

Keywords

Personnel Emotion Recognition, Feature Extraction, Transformer, Feature Fusion

1. 引言

人员情绪识别作为一个涉及到模式识别、深度学习和机器视觉等多方面的研究课题，具有广泛的应用前景和重要的研究意义。目前，人员情绪识别技术已被应用于智能教学、智慧医疗、人机交互以及 V2C 等领域中。在 V2C 中，情绪识别技术能够分析并捕捉语音中的情感特征，有助于更准确地克隆出目标语音的音色，更能够模仿出目标语音中的情感表达。

目前，人员情绪识别分类的方法主要有基于机器学习加分类器的方法与基于深度学习的方法等[1]。基于机器学习加分类器的方法虽然能够有效处理小样本数据集，但是特征提取器的性能会受到手工设计特征的影响。基于深度学习的方法主要是使用卷积神经网络与循环神经网络等模型自动从预处理后的图像中提取特征，并使用 softmax 等函数进行分类。该方法能够自动学习数据中的特征表示，无需人工干预。基于深度学习的方法具有更为强大的特征学习能力，通过使用大量数据进行训练，从而使模型的泛化能力与性能更好[2]。在深度学习领域，Ruicong ZHI 等人提出一种使用 3D 卷积神经网络提取视频序列中人脸微表情特征的方法。采用 3D 卷积神经网络进行自学习特征提取，利用迁移学习解决面部微表情数据库样本不足的问题。实验结果表明，该方法取得了较好的性能[3]。Nagaraju Melam 等人提出一种基于双优化的卷积网络模型，主要包括 U 型网络、残差网络结构和 Coot 优化。在公开视频数据集上的实验结果表明，该模型有效提高了情绪识别的准确率[4]。Sanoar Hossain 等人使用双线性池化的 CNN 架构对各种稀疏人脸数据库进行特征表示，使用面部动作单元检测识别面部活动，从而实现情绪识别。该方法能够降低计算成本，减轻网络过拟合问题，但识别精度仍有待提高[5]。A. Rajesh kumar 等人提出一种基于混合深度学习模型的视频序列面部情绪识别方法。该方法使用两个独立的深度卷积神经网络在分割的

视频片段上分别学习空间和时间特征，在进行特征融合之后送入支持向量机中进行情绪分类识别任务，公开视频数据集上的实验结果证明了该方法的有效性[6]。徐胜超等人提出一种基于多方向特征融合的动态人脸微表情识别方法。对动态图像进行去噪处理之后使用 AAM 模型提取动态人脸图像中距离特征信息，对全部特征进行多方向融合，实现动态人脸微表情识别。该方法虽取得了较好的识别效果，但并没有考虑到视频数据中的时间维度信息[7]。何晓云等人提出一种基于注意力机制的视频人脸表情识别算法。使用改进后的 VGGNet16 网络模型提取视频序列图片中的空间特征，同时使用光流法提取图片中的时间特征，并将两者进行特征融合。该方法通过将空间特征与时间特征融合的方式提高了识别准确率[8]。

尽管上述研究所提出的视频人员情绪识别方法均获得了不错的效果，但仍然存在一些问题。通过对上述研究方法的综合分析可知，在提取视频数据空间特征的同时将时间维度特征考虑在内能够有效提高模型识别准确率，增强模型性能。因此，本文提出一种结合 3D-ResNet34、ConvLSTM 以及 Transformer 模块的网络模型用于视频人员情绪识别。

2. ResNet34 网络模型

2.1. ResNet34 网络

为解决卷积神经网络随着层数不断加深而出现的梯度消失和梯度爆炸等问题，微软研究院的何恺明等人在 2015 年提出了 ResNet 这一深度残差网络[9]。ResNet34 则是 ResNet 系列中的一个经典网络结构，它的主要思想是引入残差模块，通过直接将输入的信息与输出相加，使得信息能够更加顺畅地传递[10]。这种残差连接允许跨层直接传递梯度的信息，使得网络更加深层且更容易训练。

ResNet34 的整体网络结构如表 1 所示。首先，输入图像经过一个 7 × 7 的卷积层，卷积层主要用于提取输入图像特征，通过卷积操作捕获图像中的空间信息和局部特征模式，步长为 2，产生 64 通道的特征图；其次，通过批归一化层和 ReLU 激活函数进行非线性变换；然后，通过 4 个残差块组成的堆叠层来提取更深层次的特征。每个残差块由若干个基本块或瓶颈块组成。在每个残差块内部，通过残差连接将输入与输出相加，使得信息能够直接传递[11]。最后，通过全局平均池化层将特征图转换为一维向量，并连接全连接层。全连接层输出与数据集中的类别数量相等的特征向量，使用 Softmax 层进行识别分类。

Table 1. ResNet34 overall network structure table
表 1. ResNet34 整体网络结构表

Layer Number	Layer Name	Output Size	Configuration
	Conv 1	112 × 112	7 × 7, 64, stride 2
Layer 1	Conv 2_x	56 × 56	3 × 3 max pool, stride 2 [3 × 3, 64] × 6
Layer 2	Conv 3_x	28 × 28	[3 × 3, 128] × 8
Layer 2	Conv 4_x	14 × 14	[3 × 3, 256] × 12
Layer 4	Conv 5_x	7 × 7	[3 × 3, 512] × 6
		1 × 1	Average pool, 1000-dfc, softmax

2.2. 3D-ResNet34 网络

3D-ResNet34 是一种三维残差网络模型，是对 ResNet34 模型在处理视频数据等时空序列数据时的扩展。3D-ResNet34 基于 ResNet34 模型，使用三维卷积(3D Convolution)与三维池化(3D Pooling)，与传统的

二维卷积与池化不同，三维卷积同时考虑了视频数据中的时间维度和空间维度，能够捕获视频数据中的时空信息，更适合处理视频等时空序列数据[12]，三维卷积的数学表达式如式(1)所示。

$$a_{ij}^{xyz} = \mathrm{ReLU}\left(b_{ij} + \sum_m \sum_{h=0}^{H_i-1} \sum_{b=0}^{B_i-1} \sum_{t=0}^{T_i-1} W_{ijm}^{hbt} a_{(i-1)m}^{(x+h)(y+b)(z+t)}\right) \tag{1}$$

式中，a_{ij}^{xyz} 表示第 i 层的第 j 特征图上在 (x,y,z) 位置的像素值，W_{ijm}^{hbt} 表示前一层的第 m 个特征图处于 (h,b,t) 位置的权重，H_i 代表三维卷积核的高度，B_i 代表宽度，T_i 代表时间维度，ReLU 为激活函数，b_{ij} 是卷积核的偏置。三维卷积操作的具体示意图如图 1 所示。

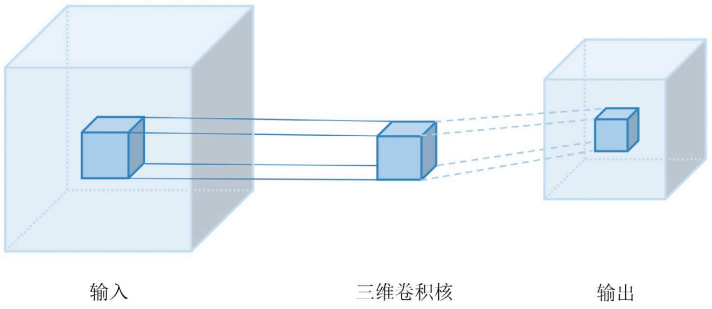

Figure 1. 3D convolution diagram
图 1. 三维卷积示意图

在三维池化中，池化操作沿着时间维度、高度维度和宽度维度进行[13]，以捕获数据立体特征。常见的三维池化方法包括三维最大池化和三维平均池化。三维池化能够减少特征图的维度，降低模型复杂度，同时保留关键的特征信息。

3D-ResNet34 网络模型同样由多个三维残差块堆叠而成，三维残差块示例如图 2 所示。每个残差块中包含若干个 3D 卷积层、3D 池化层与批量归一化层。3D-ResNet34 相比于传统的二维卷积网络，在处理视频数据时需要更多的参数和计算量。这是因为三维卷积同时考虑了时间和空间维度，使得网络能够更好地捕获视频数据中的时空信息，但增加了模型复杂度。

3D-ResNet34 主要应用于处理视频数据等时空序列数据，在视频分类、动作识别、行为分析等领域有着广泛的应用。它能够充分利用视频数据中的时空信息，提高网络对时空序列数据的建模能力和预测性能。

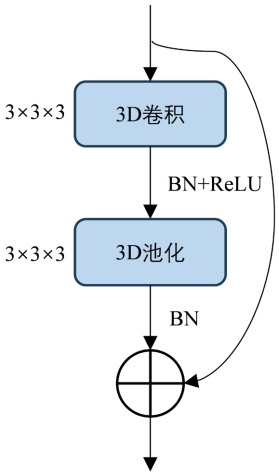

Figure 2. 3D residuals block diagram
图 2. 三维残差块示意图

3. 整体网络模型结构

本文提出了一种结合 3D-ResNet34、ConvLSTM 以及 Transformer 模块的网络模型用于视频数据中的人员心情识别。模型架构示意图如图 3 所示，具体改进包括以下几点：

(1) 将完整的视频数据划分成若干个连续的片段，每个片段包含 16 帧图像。使用 3D-ResNet34 网络对视频片段进行空间特征学习，得到每个片段的空间特征表示 $F \in R^{N\times d}$，其中 N 表示片段的数量，d 表示每个片段特征的维度，之后对片段特征进行归一化处理。

(2) 将视频片段的特征表示输入到 ConvLSTM 模块。ConvLSTM 能够同时考虑空间和时间维度的特征，使用 ConvLSTM 模块提取视频数据中的时间维度特征。

(3) Transformer 模块用于进一步处理视频数据的空间特征。通过使用多头自注意力机制让模型能够学习到每一帧图像在视频片段中的重要性或关注程度。对于每个视频片段，模型均生成一个注意力分布，用于加权融合各帧的特征，最终得到一个表示整个视频片段的注意力特征。通过注意力特征表示，模型可以更有效地捕捉视频片段中的关键信息，提高识别准确性。

(4) 对已经提取完成的时间维度特征与注意力特征进行特征融合，形成一个新的特征表示，并将融合后的特征表示送入全连接层。全连接层对输入的特征表示进行映射得到一个 1 × 7 维的向量，其中每个维度对应着一个视频片段的心情类别，向量的值为类别的预测分数，分数值最高的一类便作为视频中人员心情识别分类的最终结果。

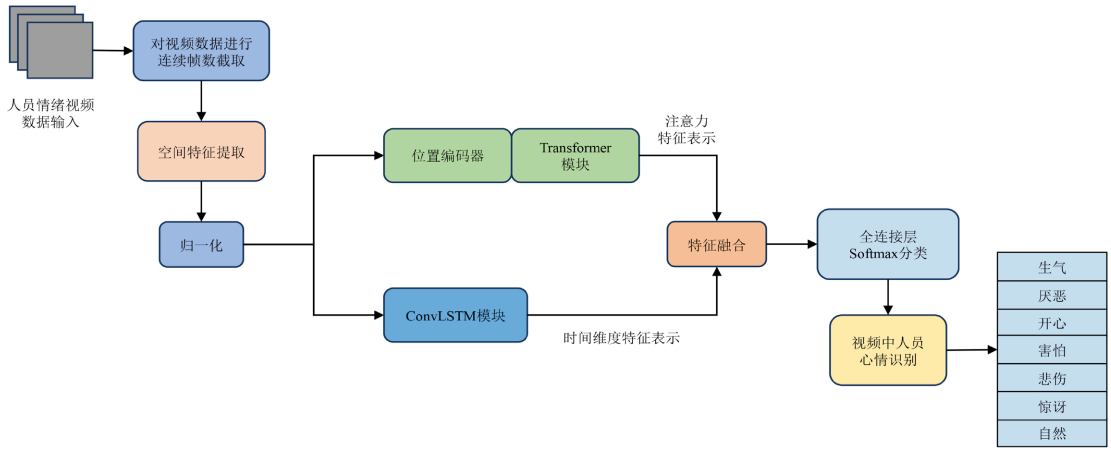

Figure 3. Overall network model structure diagram
图 3. 整体网络模型结构示意图

3.1. 人脸图像预处理

对于视频数据，首先将视频转换为帧图像形式，之后使用基于卷积神经网络的方法对图像进行人脸检测，具体模型为 MTCNN 模型。MTCNN 分别由 P-Net (Proposal Network)，R-Net (Refine Network)和 O-Net (Output Network)三部分组成，其中 P-Net 负责生成候选人脸框，即对可能包含人脸的区域进行快速筛选；R-Net 负责在候选框的基础上进行更准确的人脸检测和定位，同时进行人脸边界框的回归和人脸与非人脸的分类；O-Net 负责进一步提高人脸检测的精度，同时执行更细粒度的关键点定位，例如眼睛、鼻子、嘴巴等。针对完整视频数据过长的问题，本文采用固定间隔采样法对完整视频数据进行处理。以一个固定的时间间隔来选取视频数据中的连续帧作为采样帧，连续的帧数定为 16 帧，根据每个视频的完整帧数来确定时间间隔。该方法能够简化数据处理流程，降低计算成本。

3.2. 特征提取

(1) 空间特征提取

本文采用改进之后的 3D-ResNet34 网络提取空间特征，空间特征主要包括人脸位置、面部表情变化以及眼睛、嘴巴等部位的形态信息等。首先将视频数据作为 3D-ResNet34 网络的输入，视频数据是一个包括时间维度，通道维度，高度维度以及宽度维度的四维张量。然后在 3D-ResNet34 网络中使用 3D 卷积操作对视频数据进行特征提取。3D-ResNet34 网络中的每个残差块内部都包含多个 3D 卷积层、批量归一化层以及残差连接。在经过多个残差块的堆叠后，3D-ResNet34 网络能够从视频数据中提取出一系列的特征映射，特征映射对应着视频中不同层次的空间特征。最终将特征映射汇总，形成视频数据中的空间特征表示。

(2) 时间维度特征提取

因视频数据中的时间维度信息能够让网络模型更好地理解和分析人的心情变化模式，提高心情识别的准确性。所以采用 ConvLSTM 模块从空间特征信息表示中提取更深层次的时间维度特征。

输入数据包括时间序列数据，每个时间步的输入数据维度为(T,H,W,C)，其中 T 是时间步数，H 和 W 是输入数据的高度和宽度，C 是输入数据的通道数。在每个时间步 t 上，对输入数据进行卷积操作，提取空间特征，得到输出。将卷积操作得到的输出数据输入到 LSTM 单元中进行计算。给定 LSTM 单元有 N 个隐藏单元，输入数据的时间步数为 T，则 LSTM 模块中的细胞单元处理序列数据的过程表示如下：

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \tag{3}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{4}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o\right) \tag{5}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{6}$$

其中，x_t 是卷积操作的输出数据，h_t 是当前时间步的隐藏状态，c_t 是当前时间步的细胞状态，i_t、f_t、o_t 分别是输入门、遗忘门和输出门，W 和 b 是模型参数，σ 是 Sigmoid 激活函数，tanh 是双曲正切函数。将 LSTM 单元的输出 h_t 作为当前时间步的输出，用于下一个时间步的计算。

(3) Transformer 模块设计与注意力特征提取

本文在网络模型中设计添加的 Transformer 模块由一个 Multi-Head Attention 层、位置编码器以及一个 FeedForward 层组成，具体模块结构与特征提取流程如图 4、图 5 所示。FeedForward 层中包含两个全连接层和一个 Leaky ReLU 激活函数，用于在 Transformer 模块内部对输入进行非线性映射和特征变换。Transformer 模块的输出则是一个表示整个视频片段的注意力特征。具体如下：

$$X = C_{concat}^{(v)}\left(X_{class}, X_1, \cdots, X_N\right) \tag{7}$$

$$X^l = \mathrm{MHA}\left(\mathrm{LP}\left(X^{l-1}\right)\right) + X^{l-1}, l \in [1, L] \tag{8}$$

$$Z_{class} = \mathrm{BN}\left(\mathrm{MLP}\left(X^L(0)\right) + X^L(0)\right) \tag{9}$$

其中，MHA 为多头自注意力机制，LP 表示线性映射，BN 是归一化，X^{l-1} 和 X^l 表示前一层的注意力矩阵

和当前层的注意力矩阵，Z_{class} 是最后输出的注意力特征。

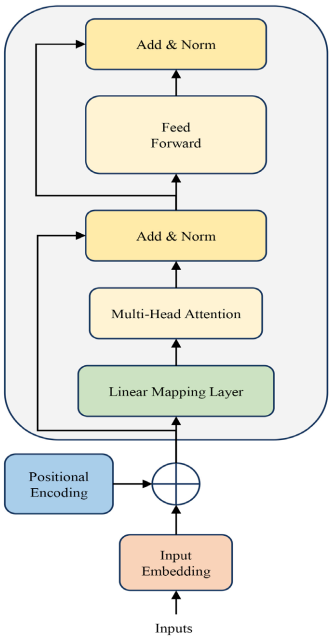

Figure 4. Transformer module diagram
图 4. Transformer 模块图

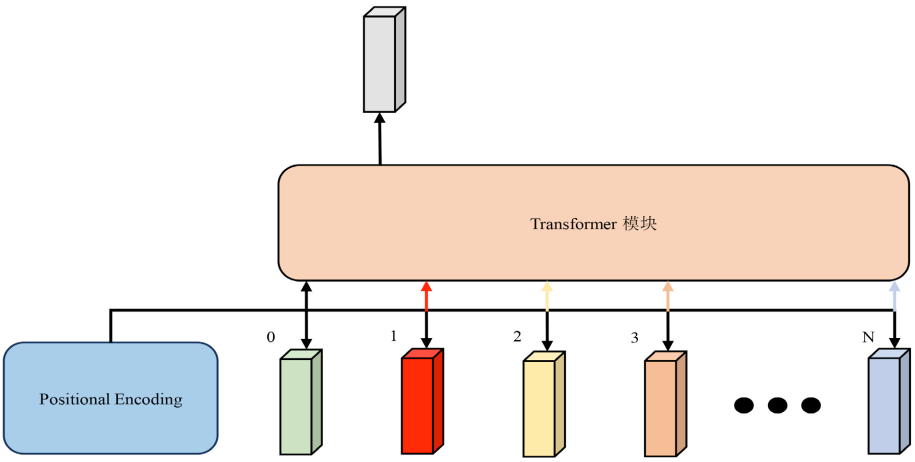

Figure 5. Diagram of feature extraction process of Transformer module
图 5. Transformer 模块特征提取过程示意图

3.3. 人员心情分类

首先对已经提取完成的时间维度特征与注意力特征进行特征融合，形成一个新的特征表示，并将融合后的特征表示送入全连接层。全连接层对输入的特征表示进行映射得到一个 1 × 7 维的向量，向量值表示类别的预测分数。全连接层的输出直接输入到 Softmax 层中进行最终的视频中人员心情分类。

4. 实验结果分析

4.1. 实验平台及相关参数设置

本文所有的对比实验与消融实验均在同一平台下运行，具体如下：

硬件方面：CPU 采用 AMD Ryzen 7 6800H (基础频率为 3.2 GHz)，GPU 采用 NVIDIA GeForce RTX3060 (16 GB DDR5)，硬盘为 512 GB SSD。

软件方面：操作系统为 Windows11 家庭版，开发环境使用 PyCharm2022，Python 版本为 3.9，深度学习框架为 Tensorflow2.10，CUDA 版本为 11.7。

本文设计的网络模型在训练时，Batch_Size (批大小)为 4，初始学习率为 1×10^{-3}，在训练过程中使用自适应矩估计优化器(Adam)进行自适应地调整学习率。Epoch 为 100。软硬件各参数如上。使用交叉验证方法，将 Oulu-CASIA 视频数据集平均分成 10 组，轮流将其中的 9 组作为训练集，剩余的 1 组作为测试集，最终的识别准确率为各测试集准确率的平均值。

4.2. 数据集介绍

本文主要在 Oulu-CASIA 视频数据集上进行验证对比。Oulu-CASIA 数据集是一个用于人员心情识别研究的公开数据集，由芬兰奥卢大学和中国科学院自动化研究所合作创建。Oulu-CASIA 数据集包含来自 80 名志愿者的 4800 个视频片段，涵盖六种基本人脸表情，包括高兴、难过、厌恶、惊讶、害怕和生气。数据集中的每个视频片段都经过了专业标注，标注内容包括人脸位置、情感类别等。为模拟真实生活中的各种情况，数据集中的视频片段是在不同的光照条件、面部遮挡情况以及姿势变化下进行拍摄的[14]。具体示例如图 6 所示。

Figure 6. Oulu-CASIA face expression image example
图 6. Oulu-CASIA 人脸表情图像示例图

4.3. 评价标准

本研究采用准确率(Accuracy)作为评价指标来评估模型整体精度。利用测试集验证模型性能，并生成混淆矩阵(Confusion Matrix)，比较分类结果与实际测得值之间的差异，进而分析神经网络模型优缺点。

准确率计算公式如下所示：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

其中，TP、TN、FP 和 FN 分别是真正类(True Positive)、真负类(True Negative)、假正类(False Positive)和假负类(False Negative)。

4.4. 对比实验与分析

为进一步测试改进之后的 3D-ResNet34 网络模型性能，设计对比实验，与 VGG16 网络模型、3D-ResNet34 基础网络模型等以及相关文献中改进后的模型进行比较，具体结果如表 2 所示。

Table 2. Accuracy comparison of Oulu-CASIA dataset on different models
表 2. Oulu-CASIA 数据集在不同模型上的准确率对比

方法	准确率/%
3D-Resnet34	84.05
VGG16	71.59
ALFWNet [15]	85.42
PHRNN-MSCNN [16]	86.25
LBVCNN [17]	82.41
本文方法	88.37

由表 2 中的数据可以看出，本文改进之后的 3D-ResNet34 结合 ConvLSTM 以及 Transformer 的网络模型与原基础网络相比，识别准确率提高了 4.32%。同时，与其他的经典网络模型以及相关文献中改进后的模型相比，本模型的识别准确率均高于其他模型，证明本文提出的网络模型具有较好的性能。

4.5. 消融实验与分析

本文模型主要由 3D-ResNet34、ConvLSTM 和 Transformer 三个模块组成，为测试模型中各个改进部分对最终识别准确率的影响程度以及各模块结构的有效性，在 Oulu-CASIA 数据集上进行消融实验，实验结果如表 3 所示。

Table 3. Ablation experiments on Oulu-CASIA dataset
表 3. Oulu-CASIA 数据集上消融实验

3D-Resnet34	ConvLSTM	Transformer	准确率/%
√	-	-	84.05
√	√	-	87.19
√	-	√	86.72
√	√	√	88.37

表 3 中打勾表示该模块已被添加到网络模型中，未打勾则表示模型暂不添加此模块，根据准确率判断各个改进模块的有效性。通过观察表中的准确率数据可知，改进之后的单独 3D-ResNet34 网络在 Oulu-CASIA 数据集上的准确率为 84.05%，引入 ConvLSTM 模块之后，网络能够学习到视频片段中的时间维度特征信息，因此准确率上升了 3.14%；在 3D-ResNet34 网络中仅添加 Transformer 模块的准确率为 86.72%，上升了 2.67%；将改进后的 ConvLSTM 模块与 Transformer 模块同时加入到 3D-ResNet34 网络中，整体网络的识别准确率达到最高，为 88.37%。实验结果表明，改进后的各模型中均具有效性，同时在网络模型中添加各模块能够大幅提高网络性能，识别准确率更高。

5. 结语

在对视频数据进行模型训练时，为综合考虑视频序列中的时间维度信息，本文提出一种结合 3D-ResNet34、ConvLSTM 以及 Transformer 模块的网络模型用于人员情绪识别。首先，将完整的视频数

据划分成若干个连续的片段，每个片段包含 16 帧图像。使用 3D-ResNet34 对视频片段进行空间特征提取。其次，在网络模型中设计添加 ConvLSTM 模块用于从空间特征中提取更深层次的时间维度特征，将视频数据中的时序信息考虑在内。然后，使用自设计的 Transformer 模块让网络能够获取到每一帧图像在视频片段中的重要性，对每个视频片段生成一个注意力分布，用于加权融合各帧的特征，得到注意力特征。最后，对提取完成的时间维度特征与注意力特征进行特征融合，并送入 Softmax 层进行分类识别。实验表明，本文设计的方法取得了较好的识别准确率。

参考文献

[1] Li, S. and Deng, W.H. (2020) Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, **13**, 1195-1215.

[2] 唐宏, 向俊玲, 陈海涛. 多区域融合轻量级人脸表情识别网络[J]. 激光与光电子学进展, 2023, 60(6): 81-89.

[3] Zhi, R., Xu, H., Wan, M. and Li, T. (2019) Combining 3D Convolutional Neural Networks with Transfer Learning by Supervised Pre-Training for Facial Micro-Expression Recognition. *IEICE Transactions on Information and Systems*, **102**, 1054-1064. https://doi.org/10.1587/transinf.2018edp7153

[4] Nagaraju, M., Yannam, A., Sreedhar P, S.S. and Bhargavi, M. (2022) Double Optconnet Architecture Based Facial Expression Recognition in Video Processing. *The Imaging Science Journal*, **70**, 46-60. https://doi.org/10.1080/13682199.2022.2163344

[5] Hossain, S., Umer, S., Rout, R.K. and Tanveer, M. (2023) Fine-Grained Image Analysis for Facial Expression Recognition Using Deep Convolutional Neural Networks with Bilinear Pooling. *Applied Soft Computing*, **134**, Article ID: 109997. https://doi.org/10.1016/j.asoc.2023.109997

[6] Kumar, A.R. and Divya, G. (2020) Learning Effective Video Features for Facial Expression Recognition via Hybrid Deep Learning. *International Journal of Recent Technology and Engineering* (*IJRTE*), **8**, 5602-5604. https://doi.org/10.35940/ijrte.e6767.018520

[7] 徐胜超, 叶力洪. 基于多方向特征融合的动态人脸微表情识别方法[J]. 计算机与数字工程, 2022, 50(8): 1818-1822.

[8] 何晓云, 许江淳, 史鹏坤. 基于注意力机制的视频人脸表情识别[J]. 信息技术, 2020, 44(2): 103-107.

[9] Li, D.H., *et al*. (2022) Emotion Recognition of Subjects with Hearing Impairment Based on Fusion of Facial Expression and EEG Topographic Map. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, **31**, 437-445.

[10] 史志博, 谭志. 融入注意力的残差网络表情识别方法[J]. 计算机应用与软件, 2023, 40(9): 222-228.

[11] Qu, Z. and Niu, D. (2023) Leveraging ResNet and Label Distribution in Advanced Intelligent Systems for Facial Expression Recognition. *Mathematical Biosciences and Engineering*, **20**, 11101-11115. https://doi.org/10.3934/mbe.2023491

[12] Pan, H., Xie, L. and Wang, Z. (2023) C3DBed: Facial Micro-Expression Recognition with Three-Dimensional Convolutional Neural Network Embedding in Transformer Model. *Engineering Applications of Artificial Intelligence*, **123**, Article ID: 106258. https://doi.org/10.1016/j.engappai.2023.106258

[13] Wu, C. and Guo, F. (2020) TSNN: Three-Stream Combining 2D and 3D Convolutional Neural Network for Micro-Expression Recognition. *IEEJ Transactions on Electrical and Electronic Engineering*, **16**, 98-107. https://doi.org/10.1002/tee.23272

[14] Naveen, P. (2023) Occlusion-Aware Facial Expression Recognition: A Deep Learning Approach. *Multimedia Tools and Applications*, **83**, 32895-32921. https://doi.org/10.1007/s11042-023-17013-1

[15] Xie, W.C., *et al*. (2019) Adaptive Weighting of Handcrafted Feature Losses for Facial Expression Recognition. *IEEE Transactions on Cybernetics*, **51**, 2787-2800.

[16] Zhang, K., Huang, Y., Du, Y. and Wang, L. (2017) Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *IEEE Transactions on Image Processing*, **26**, 4193-4203. https://doi.org/10.1109/tip.2017.2689999

[17] Kumawat, S., Verma, M. and Raman, S. (2019) LBVCNN: Local Binary Volume Convolutional Neural Network for Facial Expression Recognition from Image Sequences. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Long Beach, 16-17 June 2019, 207-216. https://doi.org/10.1109/cvprw.2019.00030