基于短程和长程关系的时空 Transformer 的微表情识别复现

肖航

摘要

由于微表情是自发的，因此即使试图隐藏一个人的真实情绪，也有助于推断其真实情绪。由于微表情持续时间短、强度低，其识别是情感计算中的一项艰巨任务。基于手工制作的时空特征的早期工作显示出一些希望，最近已被不同的深度学习方法所取代，这些方法现在正在争夺最先进的性能。然而，捕获局部和全球时空模式的问题仍然具有挑战性。为此，本次复现的论文提出了一种新颖的时空转换器架构——这是第一个用于微表情识别的纯粹基于转换器的方法（即不使用任何卷积网络）。该架构包括学习空间模式的空间编码器、用于时间维度分析的时间聚合器和分类头。本文对原论文进行了改进，复现了在三个广泛使用的自发微表情数据集（即 SMIC-HS、CASME II 和 SAMM）的微表情识别任务，得到了对应的 Acc 和 F1 指标。

关键词：微表情识别；长期光流；时间聚合器；自注意力机制

1 引言

面部表情在人际交流中发挥着重要作用，其识别是情感计算中最重要的任务之一。尽管对此仍存在一些分歧，但相当多的心理学家认为，尽管由于不同的文化环境，个体使用不同的语言进行交流，但他们的情感表达是相当普遍的 [11]。正确识别面部表情在一般交流中很重要，可以帮助理解人们的心理状态和情绪。

通俗地讲，“面部表情”一词在技术上更准确地称为“面部宏观表情”(MaE)。虽然面部宏观表情对于人类互动至关重要，提供了一种通用的非语言表达情感的方式 [42]，但面部宏观表情可以自愿产生，这意味着它们可以用来欺骗。换句话说，一个人的宏观表情可能并不能准确地代表他真实感受到的情感。然而，无论有意识的努力如何，感受到的情绪都会影响面部肌肉的短暂收缩，这些肌肉在心理抑制下不由自主地表达出来。由此产生的微小、突然和短暂的表情被称为微表情（ME）。在 Haggard 和 Isaacs 首次观察并认识到这是一种令人感兴趣的现象之后 [14]，然后由 Ekman 和 Friesen 报告的案例研究进一步阐述 [10]，ME 开始受到心理学家更广泛的研究，并且在过去十年吸引了计算机视觉领域的兴趣 [41]。与 MaE 相比，ME 很微妙。它们的展示时间为 0.04 秒至 0.2 秒 [11]，并且面部运动较少。这些特征使得 ME 比 MaE 更难被识别，无论是手动（即由人类）还是自动（即由计算机）。

为了同时自动捕获短程和长程关系，本文应用多头自注意力机制（MSM）而不是卷积核作为深度学习 MER 架构的基石。CNN 几乎不会学习块 1 和 N 之间的关系，但在 MSM 开

始时就已经考虑到了。基于 MSM 的网络称为 Transformer。序列元素之间的短程和长程关系可以以并行方式学习，因为 Transformer 完整地利用序列，而不是像循环网络那样顺序处理序列元素。

已发表文献中的大多数 MER 研究都是基于视频的，如 Ben 等人。详细阐述了 [3]，尽管在单帧分析方面有少量但值得注意的工作 [13,22,24]。该统计数据反映了这样的共识：为了获得最佳性能，需要考虑空间和时间信息。特别是，分别通过空间和时间特征提取和分析绝对和相对面部运动。现有的大多数手工方法都使用相同类型的算子，通过将帧视为 3D 数据来检测不同维度的空间和时间信息。由此产生的具有统一格式的时空特征被一起用于实现基于视频的 MER。在基于深度学习的方法中，主要通过卷积神经网络来提取空间特征。一些连接从每帧提取的空间特征，另一些使用循环神经网络来导出时间信息。为了集成各种时空关系，本次复现论文的设计在空间编码器之前利用空间数据（即视频样本的每一帧）中的长期时间信息，以及时间聚合块来融合短期和长期时间之后的关系。

本次复现论文展示了如何将基于 Transformer 的深度学习架构以优于当前技术水平的方式应用于 MER。目前工作的主要贡献如下：

1）本文提出了一种用于基于视频的微表情识别的新型时空深度学习 Transformer 框架，将其命名为基于短程和长程关系的时空 Transformer （SLSTT）。本文工作是此类深度学习 MER 的第一个工作，因为它在任何阶段都没有使用 CNN，而是完全以 Transformer 架构为中心。

2) 本文使用长期光流矩阵，以特别适合 MER 的新颖方式计算，而不是原始彩色图像作为本文网络的输入。最终达到的特征结合了长期时间信息和短程和长程空间关系，并由 Transformer 编码器块导出。

3）本文设计了一个时间聚合块来连接多个 Transformer 编码器层从每帧提取的空间关系的时空特征，并实现基于视频的 MER。还介绍了平均值和 LSTM（长短期记忆）聚合器的实证性能和分析。

2 相关工作

2.1 微表情识别

自 2013 年 SMIC 数据集发布以来，自发微表情识别的研究量逐年稳步增长。从早年的手工计算机视觉方法到最近的深度学习方法，微表情特征提取的主要思想可以归类为主要追求空间策略或时间策略。

2.1.1 空间特征

对于深度学习方法，CNN 模型可以被认为是两个组件的组合：特征提取部分和分类部分。卷积层和池化层执行空间特征提取。

除了基于局部外观的特征之外，还描述了用于微表情分析中的空间特征提取的许多其他策略。其中最简单和最常见的方法之一是采用面部感兴趣区域 (ROI) 分割。波利科夫斯基等人。[40] 根据面部动作编码系统（FACS）[12] 将每个面部样本分割为 12 个区域，每个区域对应一个独立的面部肌肉复合体，并对各个区域应用外观归一化。其他人修改或扩展了这一策略，例如采用不同的分割方法或不同的显着区域 - 11 [43]、16 [36]、36 [25] 而不是 Polikovsky

等人的 12。空间特征算子应用于每个 ROI 而不是整个图像，从而提供更细致的面部描述。近年来，这种策略的更原则性的等价物（因为它是学习的，而不是由人类预先确定的），可以以神经网络中应用的注意力块的形式找到，以提高其学习空间特征的能力。这些块可以为特征图生成权重掩模，帮助网络更多地关注重要区域。最近，GCN 也被用在深度学习框架中作为捕获空间信息的手段，通常使用 AU 来对应图节点。例如，雷等人。[20] 基于面部标志分割节点补丁并将它们与 AU GCN 融合。谢等人。[39] 通过全局平均池化从主干特征推断 AU 节点特征，并使用它们为 GCN 层构建 AU 关系图。这些优化措施使用先验知识（FACS 中的 AU）来增强提取的空间特征。这种网络不能直接学习远程空间关系。

2.1.2 时间特征

由于微表情最具特征的方面之一是其突然发生，因此时间特征不容忽视。虽然文献中的一些方法确实仅使用单个顶点帧而不是每个 ME 样本中的所有帧 [13,22,24,30]，但大多数方法都采用起始帧和结束帧之间的所有帧。偏移，从而以相同的基础对待该时间段内的所有时间变化。有些更进一步，采用时间帧插值（正如本文中所做的那样），以增加帧计数 [17,21,25,31,36]。

作为使用原始外观图像作为深度学习网络的输入的替代方案，一些作者提出使用光流矩阵形式的预处理数据 [19,26,38]。以这种方式，直接利用邻近时间信息。另一方面，不同的作者以各种方式研究了较长范围时间模式的学习。有些人只是通过将视频序列视为三维矩阵 [23,27,33] 来提取时间模式，而不是自然捕获单个图像的二维矩阵。其他人则采用循环神经网络 (RNN) 或 LSTM [17,18] 等结构。除了使用现成的循环深度学习策略之外，最近还出现了应用特定领域知识的方法，以使学习对于微表情分析特别有效 [38]。

2.2 计算机视觉中的 Transformers

大约十年来，卷积神经网络已经成为计算机视觉领域大多数深度学习算法的支柱。然而，卷积总是在固定大小的窗口上运行，因此无法提取远端关系。Transformer 的概念首先是在 NLP 的背景下引入的。它依赖于自注意力机制，学习序列元素之间的关系。Transformer 能够捕获序列元素之间的“长期”依赖性，这对于传统的循环模型编码来说是一个挑战。通过将图像划分为子图像并对它们施加一致的排序，可以将平面图像转换为序列，因此可以以与时间特征相同的方式学习空间依赖性。因此，基于 Transformer 的深度学习架构最近引起了计算机视觉社区的极大关注，并开始在许多计算机视觉任务中发挥越来越重要的作用。

目标检测背景下的一个代表性示例是 DEtection TRansformer (DETR) [4] 框架，它首先使用 Transformer 块进行回归和分类，但视觉特征仍然由基于 CNN 的主干提取。Chen 等人的图像生成预训练（iGPT）方法。[6] 尝试以不同的方式利用 Transformer 的优势，预训练 BERT（来自 Transformers 的双向编码器表示）[8]，最初是为了语言理解而提出的，然后用一个小的分类头对网络进行微调。iGPT 在 BERT 中使用像素代替语言标记，但由于必要的图像分辨率降低而导致严重的信息丢失。在分类方面，Dosovitskiy 等人的 Vision Transformer (ViT) 方法。[9] 应用图像块的 Transformer 编码作为直接提取视觉特征的手段。它是第一个纯视觉 Transformer，其精神和设计忠实地遵循了原始 Transformer [34] 架构。因此，它可以毫不费力地促进 NLP 中使用的可扩展 Transformer 架构的应用。

继这些成功之后，Transformer 已应用于各种计算机视觉任务，包括情感计算领域的任务 [5,37]。值得注意的例子包括面部动作单元检测 [16] 和基于面部图像的宏表情识别 [29]。然

而，由于本文在前面几节中讨论的挑战带来的设计困难，现有的微表情识别方法都没有充分利用空间和时间信息。

3 本文方法

3.1 本文方法概述

在目前的工作中,本次复现论文提出了一种利用微表情及其特征的生物学理解以及 Transformer 框架的方法。该方法克服了上一节讨论的文献中现有 MER 方法的许多弱点。重要的是，本文方法能够提取近端（即短程）和远端（即长程）时空特征，从而受益。本文所提出框架如图 1 所示：

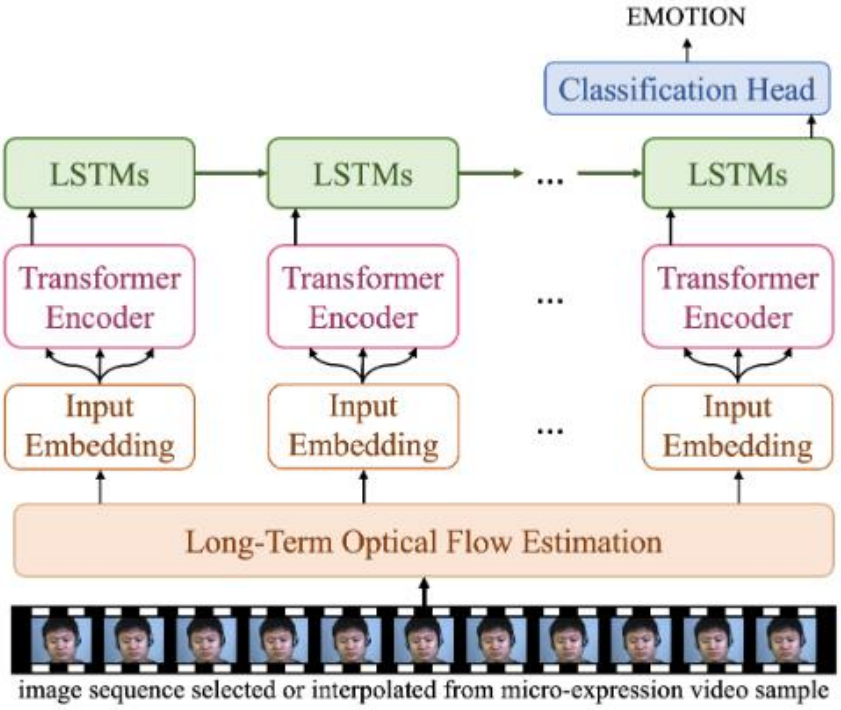

图 1. 模型框架

3.2 长期光流

光流描述了帧之间亮度模式的表观运动，这是由场景内容和用于成像的相机的相对运动引起的 [32]。如果相机是静态的，则可以使用光流从帧之间像素外观的变化来推断成像对象运动的方向和幅度 [1]。

光流本质上是时间局部的，即除了实际考虑（数值、效率等）之外，它是在连续的序列帧之间计算的。当考虑微表情视频时，这会带来一个问题，微表情视频是由表情期间表现出的有限运动创建的。因此，在这里本文提出计算每个样本帧和起始帧之间的光流，而不是计算连续帧之间的光流，见图 2。要了解这种选择背后的原因，请考虑图 3，它显示了从微表情起始帧。可以容易地观察到，直到顶点框架，场都相当相似，这可以归因于上述表达式的简洁性，此后具有相似的趋势，但方向相反。相比之下，本文的时间非局部修改光流（可以说是长

期光流）表现出一种更加结构化的模式，始终处于同一方向，直到顶点帧幅度增加，幅度减小此后。这导致与每个微表情相关的更加稳定和有区别的特征。

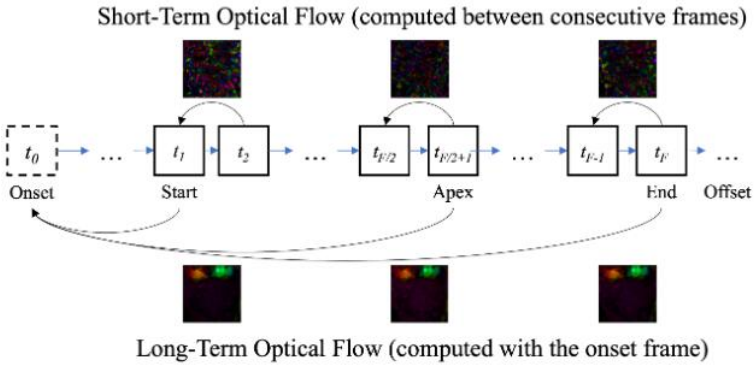

图 2. 短期和长期光流的不同计算机制

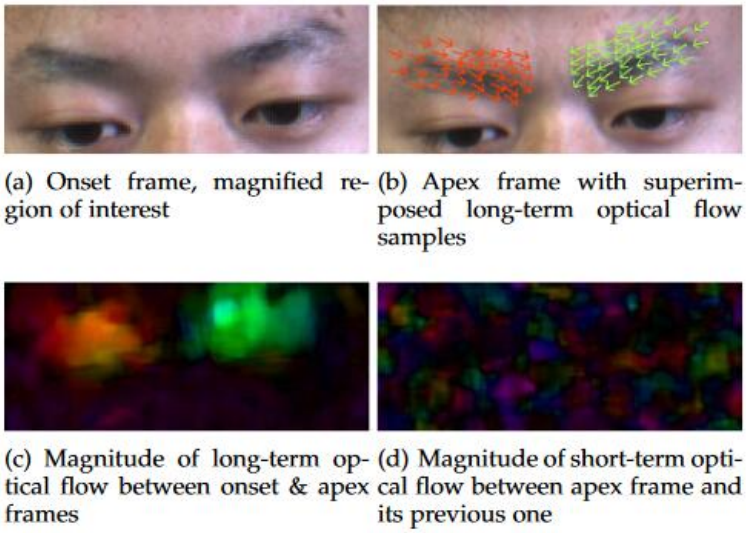

图 3. 起始帧和顶点帧之间计算的光流图示

3.3 空间特征提取

该方法的关键思想在于使用 Transformer 编码器从每个帧中提取远程空间关系，图像像以前一样被视为组成块的序列。更具体地说，输入帧首先被表示为具有每个图像块的局部空间特征的向量序列。然后将所得序列输入 Transformer 编码器以进行长期空间特征提取。

3.3.1 输入嵌入和短程空间关系学习

标准 Transformer 接收一维序列作为输入。为了处理 2D 图像，本文将每个图像表示为一系列光栅化的 2D 块。这里本文不使用外观图像（即原始视频序列帧）作为输入，而是使用相应的光流场。提出了输入嵌入块作为将输入图像表示为向量序列以输入到 Transformer 编码器的方法。

一般的输入嵌入机制将图像 $X \in \mathbb{R}^{H\times W\times C}$ 视为不重叠的 $P \times P$ 像素块序列，其中 H、W 和 C 分别是输入的高度、宽度和通道数。与 Dosovitskiy 等人提出的“分离且平坦”的线

性补丁嵌入不同 [9]。本文首先使用逐块全连接层提取块区域中的局部空间特征。图像 X 的补丁被表示为 $X_p \in \mathbb{R}^{N\times(P^2,C)}$。如图 4 所示，本文从图像 X 中提取短程空间特征到特征图 $X \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$ ，将它们展平并转置为 N 维向量，其中 $N = \frac{HW}{P^2}$ 的结果数量每个图像中的补丁。D 维向量通过所有 Transformer 编码器层。

之后，可学习的 D 维向量与序列连接，作为类标记 $(Z_0[0] = x_{class})$，其状态作为 Transformer 编码器的输出 $(Z_{L_T}[0])$。因此，Transformer 编码器的有效输入序列长度为 N+1。然后将位置嵌入添加到序列中的每个向量。整个输入嵌入过程可以描述如下：

$$\begin{aligned} Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \ldots; X_p^N E] + E_{pos}, \\ E \in \mathbb{R}^{(P^2,C)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D}, \end{aligned} \tag{1}$$

其中 $Z_0 \in \mathbb{R}^{(N\times D)}$ 是 Transformer 编码器的输入。

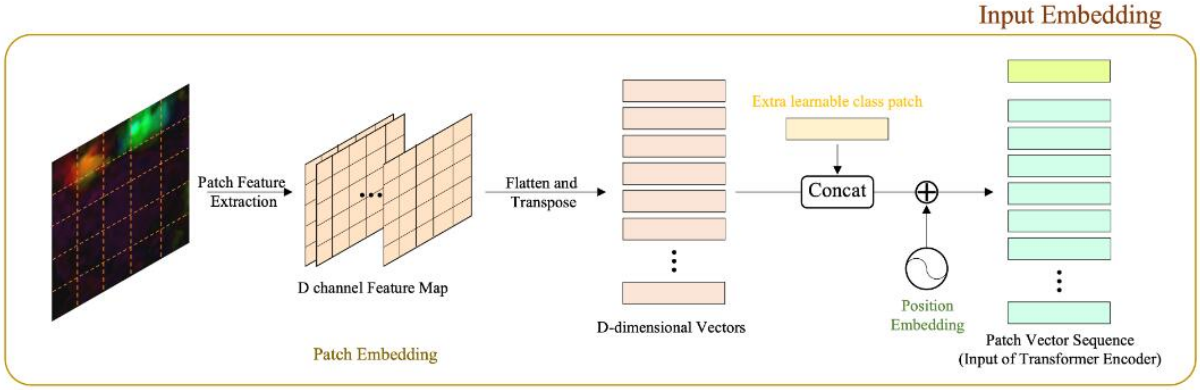

图 4. 长期光流场作为输入嵌入块的输入

3.3.2 通过 Transformer Encoder 进行远程空间关系学习

从每帧的输入长期光流场中提取短程空间关系并嵌入为向量后，它们被传递到 Transformer 编码器以进一步提取长程空间特征。本文的编码器包含 L_T 个 Transformer 层；这里使用 $L_T = 12$，该值来自 Dosovitskiy 等人的 ViT-Base 模型。[9]（本文在实验中使用的预训练编码器）。每一层涉及两个块，一个多头自注意力机制（MSM）和一个位置明智的全连接前馈网络（PWFF），如图 5 所示。在每个块之前应用层归一化（LN），在每个块之后应用残余连接 [2,35]。Transformer 层的输出可以写成如下：

$$Z_l' = MSM(LN(Z_{l-1})) + Z_{l-1}, l = 1 \ldots L_T, \tag{2}$$

$$Z_l = PWFF(LN(Z_l')) + Z_l', l = 1 \ldots L_T, \tag{3}$$

其中 Z_l 是第 l 层的输出。PWFF 块包含两个具有高斯误差线性单元 (GELU) 非线性激活函数的层。因此，特征嵌入维度首先从 D 增加到 $4D$，然后减少回 D，在本文的实验中等于 768。

多头注意力允许模型同时关注序列不同部分的信息内容，因此可以学习长程和短程空间关系。注意力函数将查询和一组键值对映射到输出（值的加权和）。权重是使用查询与相应键的兼容性函数计算的，并且它们都是向量。自注意力函数是在一组查询上同时计算的。查询、键和值可以组合在一起并表示为矩阵 Q、K 和 V ，因此输出矩阵的计算可以写为：

$$Q = Z_{l-1} W_Q, \tag{4}$$

$$K = Z_{l-1}W_K, \tag{5}$$

$$V = Z_{l-1}W_V, \tag{6}$$

$$SA(Z_l) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V, \tag{7}$$

其中 $W_Q, W_K, W_V \in \mathbb{R}^{D\times D_m}$ 是可学习矩阵，SA 是自注意力模块。MSM 可以看作是一种具有 M 个头并行操作的 self-attention 及其连接输出的投影：

$$MSM(Z_l) = \mathsf{Concat}(\{SA_h(Z_l), \forall h \in [1..M]\})W_O, \tag{8}$$

其中 $W_O \in \mathbb{R}^{M\cdot D_m\times D}$ 是重投影矩阵。D_m 通常设置为 $\frac{D}{M}$ ，以便随着 M 的变化保持参数数量恒定。

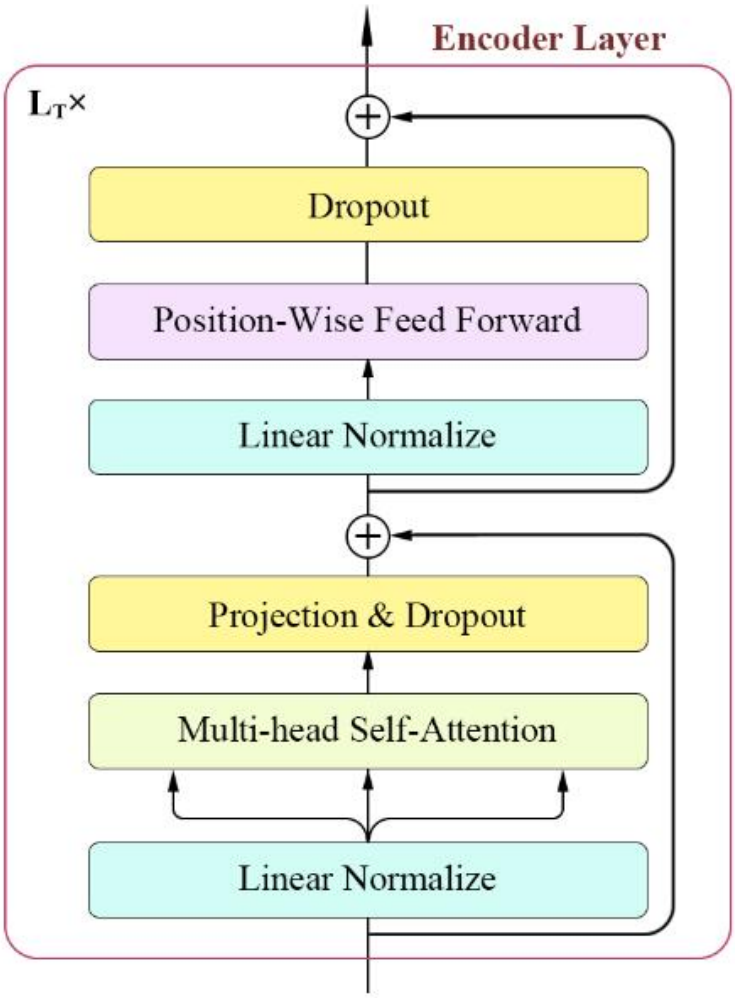

图 5. Transformer Encoder 层的详细结构

3.4 时间聚合

使用 Transformer 编码器提取与每个帧相关的局部和全局空间特征后，本文引入聚合块来在执行最终分类之前提取时间特征。聚合函数确保本文的 Transformer 模型可以被训练并应用于每个帧的空间特征集，随后处理每个样本中帧之间的时间关系。由于微表情期间的面部运动几乎难以察觉，因此单个视频样本中的所有帧都非常相似。尽管如此，仍然可以可靠

地识别许多显着框架，例如顶点框架，它们在微表情分析中起着特别重要的作用。因此，本文提出了一种用于时间聚合的 LSTM 架构。

长短期记忆（LSTM）[15] 是一种具有反馈连接的循环神经网络，它克服了与 RNN 相关的两个众所周知的问题：梯度消失问题和对时间间隔长度变化的敏感性处理序列中的显着事件。输入的元素是每帧的 Transformer 编码器的输出集。输入未连接，因此输入序列长度取决于每个 ME 视频样本中的帧数。

本文在聚合块中使用了三个 LSTM 层。每层的计算细节为：

$$\begin{aligned} t &= 1 \dots F, l = L_T + 1 \dots L_A, \\ f_t &= \sigma(W_f \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_f), \end{aligned} \tag{9}$$

$$i_l = \sigma(W_i \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_i), \tag{10}$$

$$o_t = \sigma(W_o \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_o), \tag{11}$$

$$C_t' = tanh(W_C \cdot [Z_l^{t-1}, Z_{l-1}^t] + b_C), \tag{12}$$

$$C_t = f_t \times C_{t-1} + i_t \times C_t', \tag{13}$$

$$Z_l^t = o_t \times tanh(C_t), \tag{14}$$

其中 F 是每个视频样本中所选帧的数量，L_A 是 Transformer 编码器和 LSTM 聚合器中的总层数。Z_l^t 表示处理完 t 帧后第 l 层的输出。以这种方式处理所有帧后，结果是描述整个微表情视频样本的单个特征集。最后，这些特征被输入到 MLP 中，用于最终的 MER 分类。先前的输出如何加入后面的训练的细节如图 6 所示。

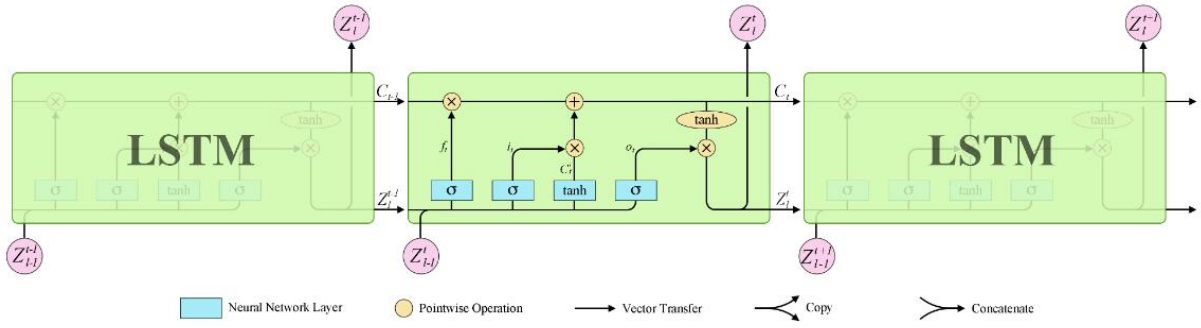

图 6. LSTM 聚合器层中的重复模块

3.5 损失函数定义

在聚合块之后，本文的网络包含两个完全连接的层，这有助于使用 SoftMax 激活函数实现最终分类。使用交叉熵损失作为训练的目标函数：

$$L = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^{C} y_{ic} \log(p_{ic}), \tag{15}$$

其中 N 是 ME 视频样本的数量，C 是情感类别的数量。当样本 i 的真实类别等于 c 时，y_{ic} 的值为 1，否则为 0。同样，p_{ic} 是样本 i 属于 c 类的预测概率。

在网络训练过程中使用梯度下降来优化目标函数时，随着参数集越来越接近其最优值，学习率应该降低。在这里，本文使用余弦退火 [28] 来实现这一点，即使用余弦函数来调节学习率，学习率最初缓慢下降，然后在再次稳定之前快速下降。这种学习率调整在当前问题的背景下尤其重要，考虑到即使在最大的语料库中，可用的微表情视频样本的数量也不是很大，如果不采取应有的谨慎，很容易学习过度拟合。

4 复现细节

4.1 与已有开源代码对比

本复现工作基于原论文开源代码：
GitHub 链接为https://github.com/Vision-Intelligence-and-Robots-Group/SLSTT。
本文对源代码中 Transformer 部分以及长期光流矩阵的计算方式进行了改动。

4.2 实验环境搭建

在空间特征提取过程中，本文采用基础 ViT 块，具有 12 个编码器层、隐藏大小为 768、MLP 大小为 3072 和 12 个头。对于初始化，本文使用在 ImageNet [7] 上预训练的官方 ViT-B/16 模型 [9]]。输入图像的大小调整为 384 × 384 像素，并将每个图像分割为 16 × 16 像素的块，以便块的数量为 24 × 24。768 维向量通过所有 Transformer 编码器层。对于时间聚合，为每个样本选择 11 个帧（顶点以及其前后的 5 个帧）作为均值聚合器和 LSTM 聚合器的输入。本文仅在实验中使用长期光流。对于学习参数，初始学习率和权重衰减分别设置为 1e-3 和 1e-4。随机梯度下降 (SGD) 的动量设置为 0.9，所有实验的批量大小为 4。所有实验均使用 PyTorch 进行，并在 Tesla P100 GPU 上进行训练。

4.3 创新点

我们对论文所提出的方法进行了改进尝试。尽管 Transformer 已经在自然语言处理领域证明了其有效性，但在视频分析和微表情识别这类时空任务中，其应用可能仍有限。运用 Transformer 的变种和改进，以更好地处理视频数据中的时空信息。长期光流矩阵可能是一个有效的特征，但也可能带来计算负担。研究更轻量级的计算方法或优化现有的光流矩阵计算方法，以减少计算时间和资源消耗。

5 实验结果分析

本文实验如表1所示，从表中可以看出，分别在 SMIC-HS、CASME II 和 SAMM 数据集上进行了测试，复现结果基本上达到了原论文的效果。

6 总结与展望

本文提出了一种基于 Transformer 的新型时空深度学习框架，用于微表情识别，这是该领域第一个完全不使用卷积神经网络的深度学习工作。在本文的框架中，可以学习样本视频的

表 1. 与原文方法的对比

	SMIC-HS		CASME II		SAMM	
	Acc(%)	F1	Acc(%)	F1	Acc(%)	F1
SLSTT(theirs)	75.00	0.740	75.81	0.753	72.39	0.640
SLSTT(ours)	73.17	0.720	71.37	0.696	67.65	0.574

空间和时间方向上的像素之间的短期和长期关系。本文使用具有多头自注意力机制的 Transformer 编码器层从可视化的长期光流帧中学习空间关系，并为时间关系设计一个时间聚合块。本文使用三个大型 MER 数据库进行实验，在单一数据库评估上取得了较好的识别精度。

本论文未来可以从下面几个方面进一步研究：

1）特征融合策略：论文中提到结合了长期时间信息和短程和长程空间关系，但没有详细描述如何融合这些特征。可以考虑采用更先进的特征融合策略来进一步提高特征提取和识别的准确性。

2）模型的可扩展性和泛化能力：为了提高模型的泛化能力，可以考虑使用更复杂的训练策略，如迁移学习、无监督学习或半监督学习。此外，为了处理不同场景和任务，模型的架构和参数也需要进一步优化。

参考文献

[1] Ognjen Arandjelović, Duc-Son Pham, and Svetha Venkatesh. Cctv scene perspective distortion estimation from low-level motion features. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(5):939–949, 2015.

[2] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. *arXiv preprint arXiv:1809.10853*, 2018.

[3] Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5826–5846, 2021.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2020.

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Paul Ekman and Wallace V Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969.

[11] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[12] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[13] Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019.

[14] Louis A Gottschalk, Arthur H Auerbach, Ernest A Haggard, and Kenneth S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. *Methods of research in psychotherapy*, pages 154–165, 1966.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[16] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7680–7689, 2021.

[17] Huai-Qian Khor, John See, Raphael Chung Wei Phan, and Weiyao Lin. Enriched long-term recurrent convolutional network for facial micro-expression recognition. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 667–674. IEEE, 2018.

[18] Dae Hoe Kim, Wissam J Baddar, and Yong Man Ro. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 382–386, 2016.

[19] Ankith Jain Rakesh Kumar and Bir Bhanu. Micro-expression classification based on landmark relations with graph attention convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1511–1520, 2021.

[20] Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1571–1580, 2021.

[21] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE transactions on affective computing*, 9(4):563–577, 2017.

[22] Yante Li, Xiaohua Huang, and Guoying Zhao. Joint local and global information learning with single apex frame detection for micro-expression recognition. *IEEE Transactions on Image Processing*, 30:249–263, 2020.

[23] Sze-Teng Liong, Yee Siang Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[24] Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018.

[25] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7(4):299–310, 2015.

[26] Yuchi Liu, Heming Du, Liang Zheng, and Tom Gedeon. A neural micro-expression recognizer. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–4. IEEE, 2019.

[27] Ling Lo, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks. In *2020 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 79–84. IEEE, 2020.

[28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[29] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021.

[30] Min Peng, Chongyang Wang, Tao Bi, Yu Shi, Xiangdong Zhou, and Tong Chen. A novel apex-time network for cross-dataset micro-expression recognition. In *2019 8th international conference on affective computing and intelligent interaction (ACII)*, pages 1–6. IEEE, 2019.

[31] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *2011 international conference on computer vision*, pages 1449–1456. IEEE, 2011.

[32] Duc-Son Pham, Ognjen Arandjelović, and Svetha Venkatesh. Detection of dynamic background due to swaying movements from motion features. *IEEE Transactions on Image Processing*, 24(1):332–344, 2014.

[33] Sai Prasanna Teja Reddy, Surya Teja Karri, Shiv Ram Dubey, and Snehasis Mukherjee. Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[36] Su-Jing Wang, Wen-Jing Yan, Guoying Zhao, Xiaolan Fu, and Chun-Guang Zhou. Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part I 13*, pages 325–338. Springer, 2015.

[37] Yiting Wang, Wei-Bang Jiang, Rui Li, and Bao-Liang Lu. Emotion transformer fusion: Complementary representation properties of eeg and eye movements on recognizing anger and surprise. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1575–1578. IEEE, 2021.

[38] Zhaoqiang Xia, Xiaopeng Hong, Xingyu Gao, Xiaoyi Feng, and Guoying Zhao. Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions. *IEEE Transactions on Multimedia*, 22(3):626–640, 2019.

[39] Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2871–2880, 2020.

[40] Liangfei Zhang and Ognjen Arandjelovic. Review of automatic micro-expression recognition in the past decade. *Machine Learning and Knowledge Extraction*, 3, 04 2021.

[41] Liangfei Zhang and Ognjen Arandjelović. Review of automatic microexpression recognition in the past decade. *Machine Learning and Knowledge Extraction*, 3(2):414–434, 2021.

[42] Liangfei Zhang, Ognjen Arandjelović, Sonia Dewar, Arlene Astell, Gayle Doherty, and Maggie Ellis. Quantification of advanced dementia patients' engagement in therapeutic sessions: An automatic video based approach using computer vision and machine learning. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5785–5788. IEEE, 2020.

[43] Liangfei Zhang, Ognjen Arandjelovic, and Xiaopeng Hong. Facial action unit detection with local key facial sub-region based multi-label classification for micro-expression analysis. In *Proceedings of the 1st Workshop on Facial Micro-Expression: Advanced Techniques for Facial Expressions Generation and Spotting*, pages 11–18, 2021.