Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(2), 185-193
Published Online May 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.132020

基于通道特征增强的火灾视频识别

丁　健，钟德军，易　云*

赣南师范大学数学与计算机科学学院，江西　赣州

收稿日期：2024年3月6日；录用日期：2024年4月9日；发布日期：2024年4月17日

摘　要

火灾对全球人民的生命财产安全造成了巨大的威胁。在火灾检测领域中，使用计算机视觉技术检测火灾对保障人民的生命和财产安全具有重要意义。针对经典的火灾识别方法无法高效地利用火焰运动特征的问题，提出基于通道特征增强的Video Swin Transformer (Video Swin Transformer based on Channel Feature Enhancement, VST-CFE)网络。VST-CFE主要包含Video Swin Transformer (VST)块和通道特征增强(Channel Feature Enhancement, CFE)块。为了利用在三维窗口划分时VST块丢失的火焰运动信息，设计了CFE块。通过建立通道信息的语义模型，CFE块增强了描述火焰运动的能力，从而提升了VST-CFE网络识别火焰的准确率。在LVFD数据集上开展大量的实验，实验结果表明VST-CFE优于基准方法VST。在该数据集上，VST-CFE的F1分数是88.16%，比基准方法VST的F1分数提高了1.75%。

关键词

通道特征增强，Transformer，火灾检测

Fire Video Recognition Based on Channel Feature Enhancement

Jian Ding, Dejun Zhong, Yun Yi*

College of Mathematics and Computer Sciences, Gannan Normal University, Ganzhou Jiangxi

Received: Mar. 6th, 2024; accepted: Apr. 9th, 2024; published: Apr. 17th, 2024

Abstract

Fires pose a huge threat to the safety of people's lives and property around the world. In the field of fire detection, the usage of computer vision technology to detect fires is of great significance for ensuring the safety of people's lives and property. Aiming at the problem that classic fire recogni-

*通讯作者。

文章引用: 丁健, 钟德军, 易云. 基于通道特征增强的火灾视频识别[J]. 人工智能与机器人研究, 2024, 13(2): 185-193.
DOI: 10.12677/airr.2024.132020

tion methods cannot efficiently utilize the motion feature of flames, a Video Swin Transformer based on Channel Feature Enhancement (VST-CFE) network is proposed. VST-CFE mainly includes the Video Swin Transformer (VST) block and the Channel Feature Enhancement (CFE) block. To utilize the motion information of flames lost in the VST block during 3D window partitioning, the CFE block is designed. By establishing the semantic model of channel information, the CFE block enhances the ability to describe flame motion, thereby improving the accuracy of the VST-CFE network in recognizing flames. Extensive experiments are conducted on the LVFD dataset, and the experimental results demonstrate that the VST-CFE method outperforms the baseline method VST. On this dataset, the F1 score of VST-CFE is 88.16%, which is 1.75% higher than the F1 score of the baseline method.

Keywords

Channel Feature Enhancement, Transformer, Fire Detection

1. 引言

近些年来，火灾在全球各地频发，对人民的生命和财产安全造成巨大的威胁。高效地检测火灾有利于保障人民的生命和财产安全。在自然语言处理领域，Transformer [1]取得了巨大的成功。在图像处理领域，Swin Transformer [2]获得了优异的成绩。在视频识别领域，Swin Transformer 的变体 Video Swin Transformer [3] (VST)有着强大的视频识别能力。VST 使用基于多头自注意力的 3D 窗口(3D Window based Multi-head Self-Attention, 3D W-MSA)使得多头自注意力(Multi-head Self-Attention, MSA)的计算集中在 3D 窗口中。该操作减少了全局 MSA 计算带来的高额计算量。VST 使用基于多头自注意力的 3D 转换窗口(3D Shifted Window based MSA, 3D SW-MSA)将窗口之间的信息关联，使得在计算量减少的同时不丢失 3D 窗口之间的关联信息。这种高效的注意力计算方式使得基于 Transformer 架构的 VST 网络能够高效地工作在通用视频识别领域。但是，随着环境、燃烧物化学性质等的变化，火灾中火焰的形状、颜色、运动状态等也会改变。VST 缺乏对火焰这种特殊物质的识别能力。

为了解决上述问题，本文提出一个基于通道特征增强的 Video Swin Transformer (Video Swin Transformer based on Channel Feature Enhancement, VST-CFE)网络。该网络主要包含 VST 块和 CFE 块。在 LVFD 数据集上开展大量的实验，实验结果表明 VST-CFE 优于基准方法 VST。此外，VST-CFE 的 F1 分数是 88.16%，比基准方法 VST 的 F1 分数提高了 1.75%。本文的主要贡献如下：

1) 为了利用在三维窗口划分时 VST 块丢失的火焰运动信息，设计了 CFE 块。通过建立通道信息的语义模型，CFE 块增强了网络描述火焰运动的能力。

2) 提出基于 Swin Transformer 架构的 VST-CFE 网络来识别含有火灾的视频。在 LVFD 数据集上的实验证明 VST-CFE 优于基准方法 VST。

2. 相关工作

近些年来，从事火灾检测的研究者在火灾检测领域中探索出一系列火灾检测的方法。这些方法的提出促进了火灾检测领域的快速发展。

大多数深度学习模型必须在性能和检测准确率之间进行平衡，以维持合理的推理时间和参数量。针对该问题，Jadon 等人[4]提出名为 FireNet 的“从头开始设计”的轻量级、更好性能的神经网络。Shees 等人[5]对 FireNet 进行改进，提出了适用于早期火灾检测的轻量级卷积神经网络。图像或视频中检测火焰对于早期火灾预警系统非常重要。针对该问题，Aliser 等人[6]提出使用注意力模块的深度网络架构对火焰进行分割检测。森林环境复杂，森林中的烟雾类物体常常干扰烟雾识别。边缘烟雾浓度稀薄，容易导致边缘遗漏。针对这些问题，Li 等人[7]提出了一种高精度边缘聚焦森林火灾烟雾检测网络。现有的深度学习模型很难平衡准确性和轻量级设计。针对这一问题，Jin 等人[8]提出一种新的轻量级深度学习算法。早期火灾的火焰很小，传统的火灾探测器无法有效探测到。针对这一问题，受火焰颜色特征的启发，Li 等人[9]提出了浅引导深度网络来解决现有早期火灾检测模型中的问题。基于视频的火灾探测模型严重依赖标记数据，并且数据标记过程特别昂贵且耗时。针对该问题，Lin 等人[10]提出了半监督火灾检测模型。由于结构的复杂性，目前基于 DETR 的火灾探测模型需要大量的内存和较长的推理时间，实用性较差。同时，高质量的火灾检测数据集非常稀缺，严重限制了算法的性能。针对这些问题，Zheng 等人[11]提出了基于扩散模型的数据集质量增强框架，以提高低质量火灾报警数据集的质量。针对 YOLO 系列模型的不足，Liu 等人[12]提出了基于注意力增强幻影模式、混合卷积金字塔和火焰中心检测的 YOLO 火灾检测算法。目前几乎没有可用于机器学习的，可学习的早期火灾数据集。针对这一问题，Kim 等人[13]提出了针对某些空间进行优化的早期火灾探测系统。该系统针对每个空间使用基于数字孪生的自动火灾学习数据生成模型。针对传统卷积神经网络在不同森林环境中固有的局限性，El-Madafri 等人[14]提出新颖的分层域自适应学习框架，旨在增强野火检测能力。该框架创新性地采用了双数据集方法，集成非森林和森林特定数据集来训练善于处理不同野火场景的模型。针对模型在大规模火灾区域和复杂森林环境的背景下的特征表示能力和检测精度的不足，Xu 等人[15]将 ConvNeXtV2 [16]引入到 YOLOv7 [17]算法中，结合多种注意力模块提高了模型对火灾的检测性能。古建筑中火灾检测需要快速、准确和实时。针对该问题，陈等人[18]对 FireNet 进行改进，提出一个新的模型。该模型利用注意力机制与多尺度特征融合实现了火灾检测能力的提高。

综上所述，在前面所提到的方法中，大部分方法基于图像处理技术来检测火灾，没有使用火焰的时序信息。这容易把外观与火焰相似并且运动特征与火焰不同的物体检测为火焰。针对上述问题，设计 CFE 块。通过建立通道信息的语义模型，CFE 块增强了描述火焰运动的能力，从而提升了 VST-CFE 网络识别火焰的准确率。

3. VST-CFE 模型

为了高效地利用火焰运动特征，提出基于通道特征增强的网络 VST-CFE。如图 1 所示，其主要包含 VST 块、CFE 块、预测块等。

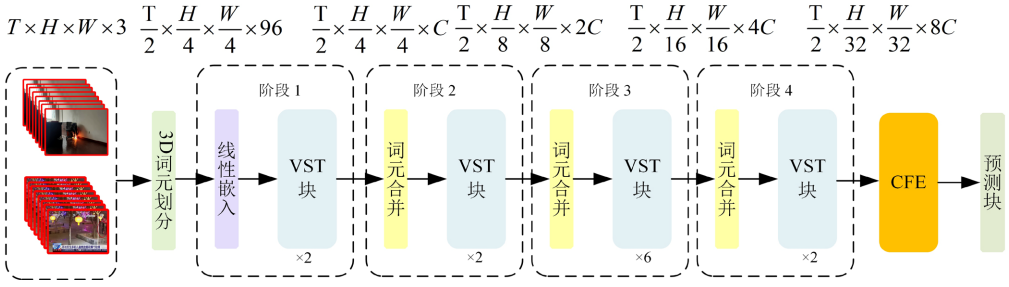

Figure 1. Schematic diagram of VST-CFE
图 1. VST-CFE 架构示意图

3.1. VST 块

VST-CFE 网络中所使用的核心模块之一是 VST 块。在 VST-CFE 网络中，VST 块成对存在。在第一个 VST 块中，使用层归一化(Layer Normalization, LN)对输入特征进行特征量级统一。处理后的特征经过 3D W-MSA 模块。在该模块中对输入的 3D 词元在 3D 窗口中进行 MSA 计算。若输入视频的 3D 词元的个数为$T'\times H'\times W'=8\times 8\times 8$且设置的 3D 窗口的大小为$P\times M\times M=4\times 4\times 4$，则一个视频由 8 个非重叠 3D 窗口组成。将输出的特征通过 LN 处理后通过多层感知机(Multi Layer Perceptron, MLP)提取出 3D 窗口中较高级的语义信息。虽然第一个 VST 块在 3D 窗口中进行 MSA 计算降低了 MSA 在全局计算中所带来的损耗，但是 3D 窗口的划分丢失了 3D 窗口之间的关联信息，限制了网络对运动火焰的识别能力。

第二个 VST 块所采用的 3D SW-MSA 增加了第一个 VST 块丢失的 3D 窗口间的联系信息。第二个 VST 块的输入特征为第一个 VST 块的输出特征。第二个 VST 块的整体结构与第一个 VST 块相似。为了增加 3D 窗口之间的关联信息，3D 窗口沿着T'，H'，W'的方向转移(2, 2, 2)个 3D 词元。转移后形成$3\times 3\times 3=27$个不同大小的 3D 窗口。为了实现在不增加 3D 窗口数量的情况下实现批量运算，采用文献[2]中提到的策略来进行批量运算。实现方法是将较小的 3D 窗口组合成大小为$P\times M\times M=4\times 4\times 4$的窗口。原来在$T'$，$H'$，$W'$轴上不相邻的 3D 词元不进行注意力计算，从而形成 8 个相同大小的 3D 窗口。该方法实现了在不增加批量个数的情况下批量运算。

MSA 一个头的注意力计算如式(1)所示。在实验中设置第一个阶段采用 3 个注意力头，第二个阶段采用 6 个注意力头，第三个阶段采用 12 个注意力头，第四个阶段采用 24 个注意力头。

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{d}}+\mathbf{B}\right)\mathbf{V} \tag{1}$$

其中，**Q** 表示查询，**K** 表示键，**V** 表示值，d 表示特征的维度，**B** 表示偏置。两个连续的 VST 块的计算过程如式(2)所示。

$$\begin{aligned}
&\hat{\mathbf{Z}}^{l}=\text{3DW-MSA}\left(\text{LN}\left(\mathbf{Z}^{l-1}\right)\right)+\mathbf{Z}^{l-1},\\
&\mathbf{Z}^{l}=\text{FFN}\left(\text{LN}\left(\hat{\mathbf{Z}}^{l}\right)\right)+\hat{\mathbf{Z}}^{l},\\
&\hat{\mathbf{Z}}^{l+1}=\text{3DSW-MSA}(\text{LN}(\mathbf{Z}^{l}))+\mathbf{Z}^{l},\\
&\mathbf{Z}^{l+1}=\text{FFN}\left(\text{LN}\left(\hat{\mathbf{Z}}^{l+1}\right)\right)+\hat{\mathbf{Z}}^{l+1}
\end{aligned} \tag{2}$$

其中$\mathbf{Z}^{l-1}$表示输入的张量，LN 表示层归一化，3DW-MSA 表示 3D W-MSA，3DSW-MSA 表示 3DSW-MSA，FFN 表示前馈神经网络，$\mathbf{Z}^{l+1}$表示第二个 VST 块输出的张量。

3.2. CFE

CFE 是 VST-CFE 网络最重要的模块之一。通过建立通道信息的语义模型，CFE 增强了描述火焰运动的能力，从而提升了 VST-CFE 网络识别火焰的准确率。如图 2 所示。

设输入到 CFE 的张量为$\mathbf{X}\in\mathbb{R}^{T''\times H''\times W''\times C''}$，其中$T''$表示帧数，$H''$表示特征图的高，$W''$表示特征图的宽，$C''$表示通道数。CFE 提取通道特征的过程包括两个阶段。在 CFE 的第一阶段，将 **X** 通过 3D 全局平均池化形成$1\times 1\times 1\times C''$大小的张量，计算过程如下式所示。

$$\mathbf{E}=\frac{1}{T''\times H''\times W''}\sum_{k=1}^{T''}\sum_{i=1}^{H''}\sum_{j=1}^{W''}\mathbf{X}\left(k,i,j\right) \tag{3}$$

其中，**E** 表示 CFE 的第一阶段输出的张量。在 CFE 的第二阶段，将其第一阶段处理后的张量输入到两个

全连接层中，其计算过程如式(4)所示。

$$\mathbf{S}=\sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\mathbf{E}\right)\right) \tag{4}$$

其中，δ 是 ReLU 激活函数，σ 是 Sigmoid 激活函数，$\mathbf{W}_1$ 表示第一个全连接层的权重，$\mathbf{W}_2$ 表示第二个全连接层的权重。将输出的权重信息 $\mathbf{S}$ 用于增强张量 $\mathbf{X}$ 的通道特征，则定义 CFE 块的映射函数 $\mathrm{f}_{\mathrm{CFE}}\left(\mathbf{X}\right)$ 如下式所示。

$$\mathrm{f}_{\mathrm{CFE}}\left(\mathbf{X}\right)=\mathbf{S}\odot\mathbf{X} \tag{5}$$

其中，使用式(4)计算 $\mathbf{S}$，$\odot$ 是带广播机制的哈达玛积。

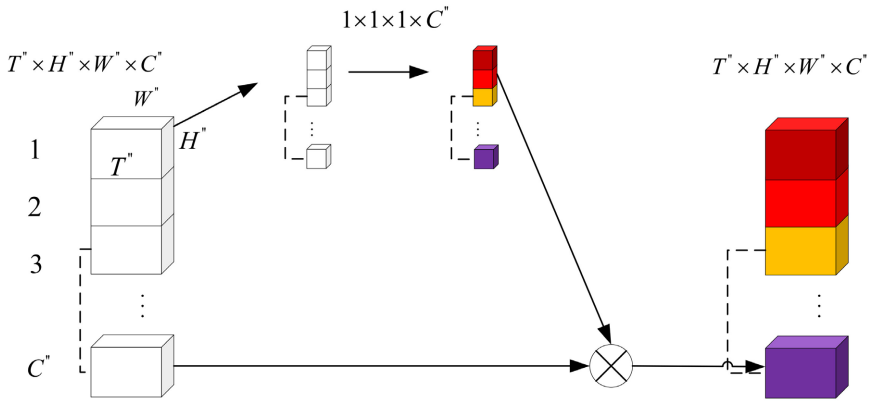

Figure 2. Schematic diagram of CFE
图 2. CFE 架构示意图

3.3. VST-CFE

VST-CFE 的网络架构如图 1 所示。假设输入 VST-CFE 网络的张量的大小为 $T\times H\times W\times 3$，其中，$T$ 表示输入视频的帧数，H 表示输入每一帧的高，W 表示输入每一帧的宽。在实验中，设置 H 为 224，W 为 224，T 为 32。如果直接将火灾视频中的帧像素作为词元，则由于词元数量巨大，导致计算复杂度过高。

为了解决这个问题，VST-CFE 网络以大小为 2 × 4 × 4 × 3 的 3D 块作为一个 3D 词元。输入的视频需要经过 3D 词元划分层。该层主要的作用是将输入的视频划分为多个大小为 2 × 4 × 4 × 3 的 3D 词元。输入的视频会形成 $\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}$ 个需要计算的 3D 词元。每个词元的特征维度为 96。线性嵌入层的作用是将特征维度变换到 96 维。

VST-CFE 网络包含 4 个特征提取阶段。第一阶段、第二阶段、第四阶段分别采用 2 个 VST 块，而第三阶段则采用 6 个 VST 块。在第一，二，三个阶段后，VST-CFE 会执行词元合并操作。这个操作对输出的特征图进行下采样。

VST-CFE 网络包含一个特征增强阶段 CFE。在该阶段中，通过修复 VST 块在 3D 窗口划分时丢失的与运动火焰相关的重要信息，CFE 增强 VST 网络对于运动火焰的识别能力。最后采用预测块来对场景视频进行分类。

3.4. 预测块

输入预测块的张量的形状为 $\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$。首先使用 3D 全局平均池化将输入的张量变化为 $1\times1\times1\times8C$。然后，该张量被调整形状为 $8C$。最后，使用全连接层将 $8C$ 投影到类别数为 2，得到预测

结果。

4. 实验

4.1. 数据集

为了验证所提出的方法，建立了火灾视频识别数据集 LVFD。该数据集包含 11,560 个视频。这些视频被分为两类，即包含火的视频和不包含火的视频。为了减少由于数据分布产生的偏置，LVFD 数据集中的视频被分为 3 组，分别是组 1、组 2 和组 3。这三个组的统计信息如表 1 所示。每组的训练集和测试集样本数的比例大概是 7:3。

Table 1. Statistics of the LVFD dataset
表 1. LVFD 数据集的统计信息

分组	类别	训练样本数	测试样本数
组 1	火	3053	1407
	无火	5099	2001
组 2	火	3197	1263
	无火	5311	1789
组 3	火	3149	1311
	无火	4984	2116

4.2. 实验规则与设置

为了更好验证本文所提出的 VST-CFE 网络对视频中运动火焰识别的高效性和鲁棒性，在 LVFD 数据集上分别进行 3 组实验。选择准确率和 F1 分数作为评估指标，其中 F1 分数是本文的主要评估指标。本文的实验步骤如下所示：

首先，为了找到最好的 F1 分数，本文在组 1 上选择训练 VST-CFE 的超参数。然后，固定训练的超参数，在组 2 和组 3 的训练集上训练网络模型。最后，在组 1、组 2 和组 3 的测试集上，分别计算实验结果。以 3 个组实验结果的平均值作为最终结果。

为了在实验过程中能更好地对本文的网络模型进行训练和测试，将数据的批量大小设置为 8，从视频中采样 32 帧。采用随机裁剪策略对数据进行增强。在训练中，较小的一侧被调整为 256 像素，从中随机裁剪 224 × 224 像素区域。以 50%的概率对每个输入帧进行水平翻转。在测试过程中，采用与训练过程类似的采样策略，将采样帧沿短边等间隔裁剪成三个区域。

在训练 VST-CFE 网络过程中，使用 AdamW 算法来优化神经网络。设置学习率为 0.0005，设置权重衰减为 0.02，用于计算梯度及其平方的运行平均值的系数分别固定为 0.9 和 0.999。此外，采用线性预热和余弦退火策略调整学习率，其中学习率从低到恒定线性增加，使用余弦退火策略衰减学习率，直到训练周期数达到 100。

4.3. 消融实验

根据上一节的实验规则，执行消融实验，验证 CFE 块在火灾视频识别领域的作用。图 3 是 VST-CFE 与基准方法 VST 在 LVFD 数据集的组 1 上的比较。其中，VST 表示没有使用 CFE 块的基准网络。

从图 3 中可以观察到，VST-CFE 和 VST 分别在第 90 和 80 个训练周期时，训练损失曲线趋于平滑，且 F1 分数获得最大值。从图 3 中的曲线可以看出，在 LVFD 数据集的组 1 上，VST-CFE 获得比 VST 更高的 F1 分数。

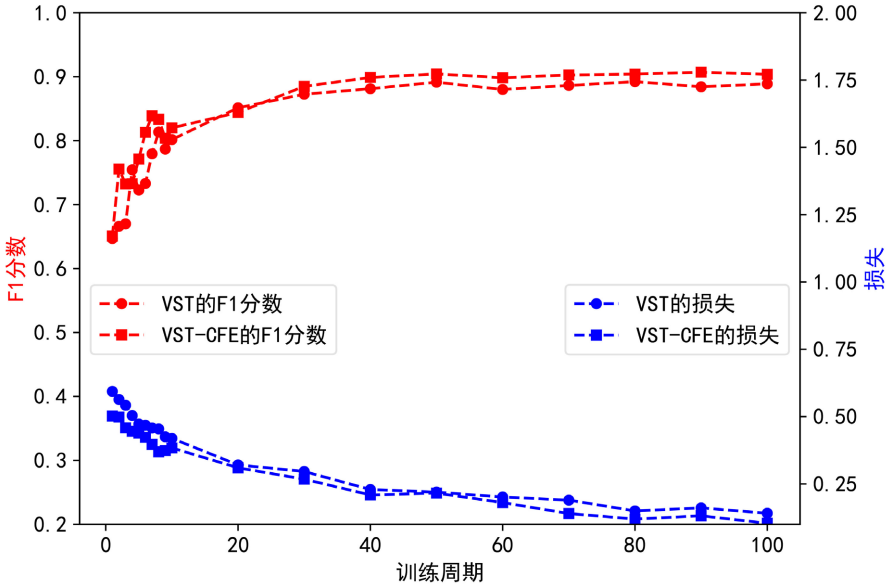

Figure 3. The comparison of VST-CFE and VST on split 1
图 3. 在组 1 上 VST-CFE 与 VST 的比较

按照上一节的实验规则，在组 2 和组 3 的训练集上，VST-CFE 网络和 VST 网络被分别训练 90 和 80 个周期。然后，在测试集上，分别计算的 F1 分数和准确率。在 LVFD 数据集上 VST-CFE 与 VST 的比较如表 2 所示。

Table 2. The comparison of VST-CFE and VST on the LVFD dataset
表 2. 在 LVFD 数据集上 VST-CFE 与 VST 的比较

分组	方法	准确率	F1 分数
组 1	VST	91.08%	89.23%
	VST-CFE	**92.46%**	**90.7%**
组 2	VST	86.93%	83.19%
	VST-CFE	**88.3%**	**85.01%**
组 3	VST	90.2%	86.82%
	VST-CFE	**91.42%**	**88.77%**

从 LVFD 数据集 3 个组的测试结果可以得出以下推论。首先，在 3 个组上，CFE 块都提升了基准网络对火焰视频识别的性能。其次，在使用 CFE 块之后，虽然在 3 个组上的 F1 分数的提升有一些差距，但是都有 1%以上的性能提升。尤其是在组 3，F1 分数的提升是 1.95%。

4.4. 与其他方法的比较

根据之前的实验规则，在 LVFD 数据集上执行实验，与经典视频识别方法 VST [3]和 TimeSformer [19]比较。VST-CFE 与其他方法的比较结果如表 3 所示。

TimeSformer 网络和 VST 网络的 F1 分数分别是 81.14%和 86.41%。这说明 VST 比 TimeSformer 更适合识别火灾视频。在加入 CFE 块之后，所提出的 VST-CFE 网络的 F1 分数达到 88.16%。这证明所提出的 CFE 块能增强与火焰运动相关的信息，从而提升网络识别火灾视频的性能。

Table 3. The comparison between VST-CFE and other methods
表 3. VST-CFE 与其他方法的比较

方法	准确率	F1 分数
TimeSformer [19]	85.67%	81.14%
VST [3]	89.40%	86.41%
VST-CFE	**90.73%**	**88.16%**

5. 总结

火焰的形状、颜色、运动等特征随着环境、燃烧物的化学性质等影响而不断变化。经典的视频识别方法缺乏描述火焰运动信息的能力。针对这个问题，本文提出 VST-CFE 网络。该网络主要包含 VST 块和 CFE 块。为了充分利用火焰的运动信息，设计 CFE 块，从而提升了 VST-CFE 网络识别火焰的准确率。在 LVFD 数据集上的实验结果表明 VST-CFE 优于基准方法 VST，并且 VST-CFE 获得最好的实验结果。

基金项目

国家自然科学基金(No. 62362003)，江西省自然科学基金(No. 20232BAB202017)，江西省研究生创新专项资金项目(YC2022-s945)。

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, December 2017, 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349

[2] Liu, Z., Lin, Y., Cao, Y., *et al.* (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

[3] Liu, Z., Ning, J., Cao, Y., *et al.* (2022) Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, 18-24 June 2022, 3202-3211. https://doi.org/10.1109/CVPR52688.2022.00320

[4] Jadon, A., Omama, M., Varshney, A., *et al.* (2019) Firenet: A Specialized Lightweight Fire & Smoke Detection Model for Real-Time Iot Applications. https://arxiv.org/abs/1905.11922

[5] Shees, A., Ansari, M.S., Varshney, A., *et al.* (2023) Firenet-V2: Improved Lightweight Fire Detection Model for Real-Time Iot Applications. *Procedia Computer Science*, **218**, 2233-2242. https://doi.org/10.1016/j.procs.2023.01.199

[6] Aliser, A. and Duranay, Z.B. (2024) Fire/Flame Detection with Attention-Based Deep Semantic Segmentation. *Iranian Journal of Science And Technology, Transactions of Electrical Engineering*, 1-13. https://doi.org/10.1007/s40998-024-00697-y

[7] Li, R., Hu, Y., Li, L., *et al.* (2024) SMWE-Gfpnnet: A High-Precision and Robust Method for Forest Fire Smoke Detection. *Knowledge-Based Systems*, **289**, Article ID: 111528. https://doi.org/10.1016/j.knosys.2024.111528

[8] Jin, L., Yu, Y., Zhou, J., *et al.* (2024) SWVR: A Lightweight Deep Learning Algorithm for Forest Fire Detection and Recognition. *Forests*, **15**, 204. https://doi.org/10.3390/f15010204

[9] Li, B., Xu, F., Li, X., *et al.* (2024) Early Stage Fire Detection System Based on Shallow Guide Deep Network. *Fire Technology*, 1-19. https://doi.org/10.1007/s10694-024-01549-1

[10] Lin, Q., Li, Z., Zeng, K., *et al.* (2024) Firematch: A Semi-Supervised Video Fire Detection Network Based on Consistency and Distribution Alignment. *Expert Systems with Applications*, **248**, Article ID: 123409. https://doi.org/10.1016/j.eswa.2024.123409

[11] Zheng, H., Wang, G., Xiao, D., *et al.* (2024) FTA-DETR: An Efficient and Precise Fire Detection Framework Based on an End-to-End Architecture Applicable to Embedded Platforms. *Expert Systems with Applications*, **248**, Article ID: 123394. https://doi.org/10.1016/j.eswa.2024.123394

[12] Liu, J., Yin, J. and Yang, Z. (2024) Fire Detection and Flame-Centre Localisation Algorithm Based on Combination of

Attention-Enhanced Ghost Mode and Mixed Convolution. *Applied Sciences*, **14**, 989. https://doi.org/10.3390/app14030989

[13] Kim, H.C., Lam, H.K., Lee, S.H., *et al.* (2024) Early Fire Detection System by Using Automatic Synthetic Dataset Generation Model Based on Digital Twins. *Applied Sciences*, **14**, 1801. https://doi.org/10.3390/app14051801

[14] El-Madafri, I., Peña, M. and Olmedo-Torre, N. (2024) Dual-Dataset Deep Learning for Improved Forest Fire Detection: A Novel Hierarchical Domain-Adaptive Learning Approach. *Mathematics*, **12**, 534. https://doi.org/10.3390/math12040534

[15] Xu, Y., Li, J., Zhang, L., *et al.* (2024) CNTCB-Yolov7: An Effective Forest Fire Detection Model Based on Convnextv2 and CBAM. *Fire*, **7**, 54. https://doi.org/10.3390/fire7020054

[16] Woo, S., Debnath, S., Hu, R., *et al.* (2023) Convnext V2: Co-Designing and Scaling Convnets with Masked Autoencoders. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, 17-24 June 2023, 16133-16142. https://doi.org/10.1109/CVPR52729.2023.01548

[17] Wang, C.Y., Bochkovskiy, A. and Liao, H.Y.M. (2023) Yolov7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, 17-24 June 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721

[18] 陈庆典, 钟晨, 刘慧, 等. 基于 Firenet 的古建筑火灾检测方法研究及改进[J]. 消防科学与技术, 2024, 43(2): 183-188.

[19] Bertasius, G., Wang, H. and Torresani, L. (2021) Is Space-Time Attention All You Need for Video Understanding? *ICML*, **2**, 4.