应用光学
Journal of Applied Optics
SCOPUS,等来源期刊
中文核心期刊、中国科技核心期刊

基于随机块移位和可变形注意力的视频烟雾识别

谢晔辉 赵海涛

Video smoke recognition based on random patch shift and deformable attention

XIE Yehui, ZHAO Haitao

引用本文:
谢晔辉, 赵海涛. 基于随机块移位和可变形注意力的视频烟雾识别[J]. 应用光学, 2024, 45(6): 1204–1211. DOI: 10.5768/JAO202445.0602005
XIE Yehui, ZHAO Haitao. Video smoke recognition based on random patch shift and deformable attention[J]. Journal of Applied Optics, 2024, 45(6): 1204–1211. DOI: 10.5768/JAO202445.0602005

在线阅读 View online: https://doi.org/10.5768/JAO202445.0602005

文章编号：1002-2082 (2024) 06-1204-08

基于随机块移位和可变形注意力的视频烟雾识别

谢晔辉，赵海涛
（华东理工大学 信息科学与工程学院，上海 200237）

摘 要：识别出工业环境中的烟雾排放行为对于规范和实时监督企业，以及环境保护都具有至关重要的意义。然而，识别工业排放烟雾具有很高的挑战性，一方面工业排放烟雾具有高透明度、高动态性等特点；另一方面烟雾的形状和尺寸可能会因环境、光照等因素而发生变化。目前主流的烟雾识别方法都是基于图像或视频的深度学习模型，但是基于图像的模型无法对视频中烟雾的动态特性进行有效的时序建模，同时基于视频的模型没有考虑烟雾形状多变的特性。将随机块移位（random patch shift, RPS）和可变形注意力（deformable attention, DA）引入 Swin Transformer。RPS 将传统的 2D 空间注意力转变为时空注意力，从而使用 2D 的自注意力计算对动态烟雾进行建模；DA 通过自适应形变的方式使网络能够适应不同的烟雾形态和外观变化，提高网络的鲁棒性和泛化能力。在 RISE 数据集上的实验结果表明，本文方法能够在 3 个子集上分别达到 0.85、0.86 和 0.84 的 F_1 分数，相比其他方法有 0.01~0.06 的提升。

关键词：烟雾识别；随机块移位；可变形注意力；深度神经网络
中图分类号：TN911.73　　**文献标志码：**A　　**DOI：**10.5768/JAO202445.0602005

Video smoke recognition based on random patch shift and deformable attention

XIE Yehui, ZHAO Haitao
（School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China）

Abstract: Recognition of smoke emission behavior in industrial environments is of vital importance for regulating and monitoring companies in real time, as well as for environmental protection. However, it is highly challenging. On the one hand, industrial emission smoke is characterized by high transparency and high dynamics, and on the other hand, the shape and size of smoke may change due to the environment, lighting, and other factors. Currently, the mainstream smoke recognition methods are deep learning models based on images and videos, but the image-based models cannot effectively model the dynamic characteristics of the smoke in the video in a time-series manner, while the video-based models do not take into account the characteristics of the variable shape of the smoke. The random patch shift (RPS) and deformable attention (DA) was introduced into the Swin Transformer. The traditional 2D spatial attention was transformed into spatio-temporal attention by RPS, thereby modeling the dynamic smoke using 2D self-attention computations. By means of adaptive deformation, DA enabled the network to adapt to different smoke shapes and appearance changes, thereby improving the robustness and generalization ability of the network. Experimental results on the RISE dataset show that the proposed method can achieve F_1 scores of 0.85, 0.86, and 0.84 in the three subsets, respectively, with an improvement of 0.01~0.06 compared to other methods.

Key words: smoke recognition; random patch shift; deformable attention; deep neural network

收稿日期：2023-09-21；　**修回日期：**2024-02-25
基金项目：国家自然科学基金（62173143）
作者简介：谢晔辉（2001—），男，硕士研究生，主要从事模式识别研究。E-mail: xie_yehui@163.com
通信作者：赵海涛（1974—），男，教授，博导，主要从事模式识别、深度学习研究。E-mail: haitaozhao@ecust.edu.cn

引言

在工业生产环境中, 烟雾的排放不仅是造成全球变暖的原因之一, 也是对人体健康造成潜在危害的危险因素之一[1]。如何准确且及时地识别出烟雾非常重要, 无论是对于安全生产还是环境保护都具有重要的意义。近年来, 基于深度学习和图像、视频处理技术的烟雾识别方法, 凭借其检测范围大、适用范围广等优点正成为新的研究热点[2]。在工业排放烟雾识别任务中, 识别的难点在于烟雾的颜色多变性, 尤其是在燃烧的早期阶段, 烟雾会呈现出白色、灰色、青色等不同颜色[3], 再加上烟雾的不透明度等特征, 这就导致烟雾难以与雾、霾、云和蒸汽等物体区别开来[4]。

现有的基于深度学习的烟雾识别模型大致可分为基于图像的方法[5-9]和基于视频的方法[10-13]。基于图像的烟雾识别方法从单帧图像中提取烟雾的静态特征, 基于视频的方法不仅从图像中学习空间信息, 也从时域中学习烟雾的动态特征。YIN Z 等人[5]设计了深度归一化卷积神经网络(deep normalization and convolutional neural network, DNCNN)来进行烟雾检测; 为了应对烟雾识别中的有雾环境, HE L 等人[6]提出了基于注意力机制的深度融合 CNN, 基于 VGG16[9]的 CNN 架构结合了空间注意力和通道注意力机制; 为了进一步提取烟雾的特征, LIU Y 等人[7]提出了一种使用暗通道先验进行图像烟雾分类的双流网络 DarkC-DCN; 随着 Transformer[14]在计算机视觉领域的发展, 基于 Transformer 的深度网络架构被用于烟雾识别; ZHOU Y 等人[8]通过使用 Vision Transformer 进行烟雾的精确检测。以上这些基于图像的深度学习模型没有考虑和利用视频的帧间信息, 因而用在视频烟雾识别问题上效果并不理想, 缺少对烟雾动态性的建模。

视频烟雾中的时序信息有时在烟雾识别中会发挥更关键的作用。考虑到视频本身是一个 3D 序列, LIN G 等人[11]提出了一种基于 Faster RCNN 和 3D CNN 的联合检测框架: 首先使用 Faster RCNN 实现静态空间信息的烟雾定位, 再用 3D CNN 结合动态时空信息实现烟雾识别。3D CNN能够明显提高精度, 但也体现出 3D CNN 参数量和计算量大的缺点。CAO Y 等人[12]利用 CNN 的中间层生成特征前景, 提出了特征前景模块(feature foreground module, FFM), 用于指导烟雾的时序建模。TAO H 等人[13]设计了一种增强扩张卷积的自适应帧选择网络 AFSNet, 用于视频烟雾识别, 该网络可以自动选取图像序列中有用的帧, 以减少信息冗余。TAO H 等人[10]还认为单凭借数据集中的二元类别标签无法有效指导网络进行烟雾识别, 因而提出了一种注意力聚合属性感知网络, 通过考虑视频属性信息(例如摄像机视角、天气等因素)来指导网络进行判别特征的学习。这些方法虽然解决了视频烟雾的动态建模问题, 但是没有考虑到烟雾的外观和纹理可能会因不同的环境和条件而变化。

为了解决以上问题, 本文提出了一种基于随机块移位和可变形注意力的 Swin Transformer[15]来进行视频烟雾识别任务。本文的研究工作主要在 3 个方面: 1) 首先将随机块移位(random patch shift, RPS)引入 Swin Transformer, 通过将空间自注意力变成稀疏的时空自注意力, 使网络能够在保持 2D 自注意力计算开销的同时, 完成对烟雾动态性的建模; 2) 可变形注意力机制[16](deformable attention, DA)的加入可以自适应地调整注意力的采样位置, 使得网络能够在对帧间信息处理的同时更好地适应不同形状和尺寸的烟雾; 3) 在 RISE 数据集上进行对比和消融实验, 最终在 3 个子集上分别达到了 0.85、0.86 和 0.84 的 F_1 分数, 相较于其他方法更具优势。

1　具有随机块移位和可变形注意力的 Swin Transformer

本文针对视频中烟雾的动态信息建模问题, 基于 2D Swin Transformer 模型, 引入随机块移位和可变形注意力机制。网络整体框架图如图 1 所示。其中, Swin Transformer Block 表示 2D Swin Transformer 中的编码器模块。在 4 层编码器中分别加入随机块移位模块, 在最后一层编码器中引入可变形注意力, 这两种模块的具体结构将分别在 1.2 与 1.3 节进行阐述。

1.1　基于视频的 Swin Transformer

输入视频 $X \in \mathbb{R}^{T\times H\times W\times 3}$, 视频一共包含 T 帧, 其中 H 和 W 分别表示输入视频帧的高和宽。采用 Video Swin Transformer[17]的 Tiny 版本作为基本框架, 将输入 X 划分为大小是 $2\times4\times4\times3$ 像素的块, 然后将维度为 $\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}$ 的张量做块嵌入(Patch Embedding), 最终把每个块的维度投影至 96, 得到网络的输入 $Z_0 \in \mathbb{R}^{\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times96}$。

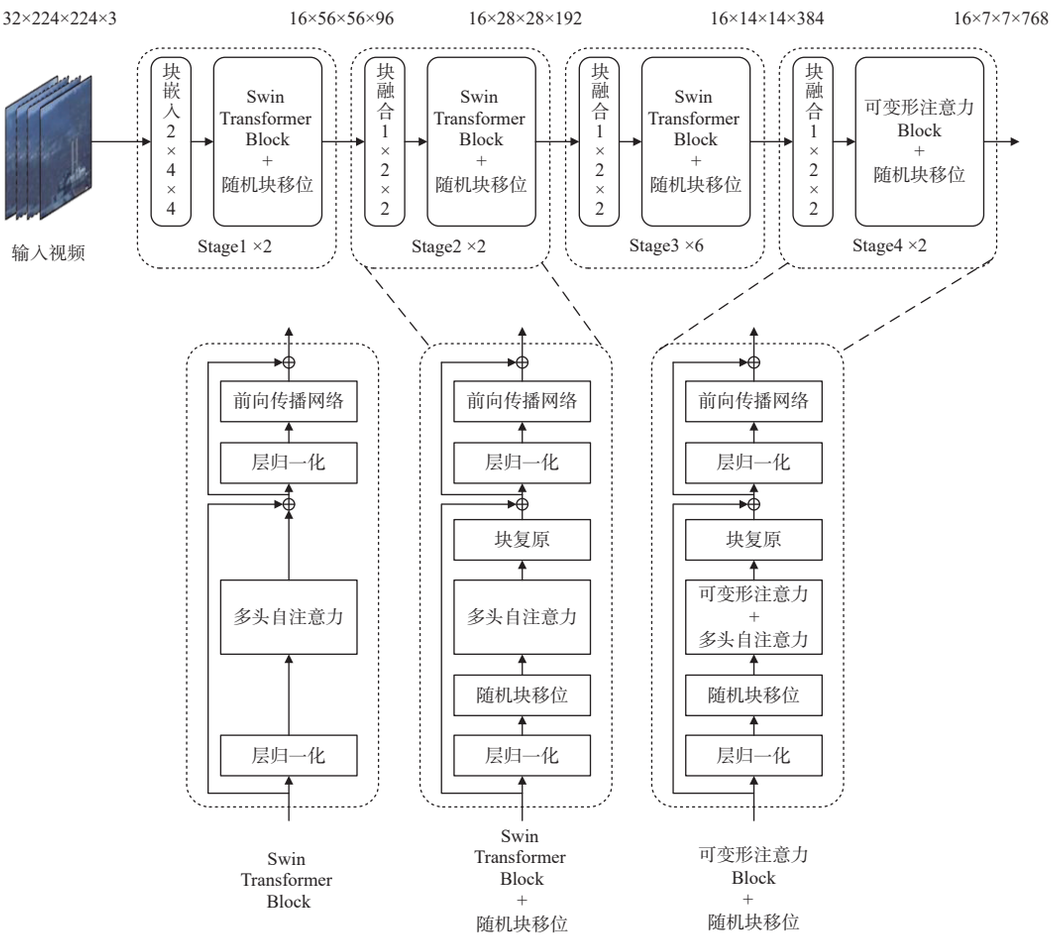


图 1 网络整体框架图

Fig. 1 Overall framework diagram of network

本文沿用 Swin Transformer 中的 2D 窗口划分机制，将输入划分为大小 $1\times M\times M$ 不重叠的 2D 窗口，M 为窗口大小，在本文中 $M=7$。网络框架中共包含 4 个编码器，每个编码器由窗口多头自注意力（window multi-head self-attention, WMSA）或移动窗口自注意力（shifted window multi-head self-attention, SWMSA）、层归一化（layer norm, LN）和前向传播网络（feed-forward networks, FFN）组成。多头自注意力操作会在每个 2D 窗口经过随机块移位操作后进行。网络中每个编码器的前向传播过程可用以下公式表示：

$$\begin{cases}\hat{Z}^{l}=\mathrm{WMSA}\left(\mathrm{LN}\left(Z^{l-1}\right)\right)+Z^{l-1}\\ Z^{l}=\mathrm{FFN}\left(\mathrm{LN}\left(\hat{Z}^{l}\right)\right)+\hat{Z}^{l}\\ \hat{Z}^{l+1}=\mathrm{SWMSA}\left(\mathrm{LN}\left(Z^{l}\right)\right)+Z^{l}\\ Z^{l+1}=\mathrm{FFN}\left(\mathrm{LN}\left(\hat{Z}^{l+1}\right)\right)+\hat{Z}^{l+1}\end{cases}\tag{1}$$

式中 $\hat{Z}^{l}$ 和 Z^{l} 分别表示多头自注意力模块和 FFN 模块输出的特征。窗口多头自注意力 SWMSA 的计算可以表示为

$$\begin{cases}\boldsymbol{Q}^{l},\boldsymbol{K}^{l},\boldsymbol{V}^{l}=W_{\boldsymbol{Q}}^{l}\mathrm{LN}\left(Z^{l-1}\right),W_{\boldsymbol{K}}^{l}\mathrm{LN}\left(Z^{l-1}\right),W_{\boldsymbol{V}}^{l}\mathrm{LN}\left(Z^{l-1}\right)\\ \mathrm{Attention}\left(\boldsymbol{Q}^{l},\boldsymbol{K}^{l},\boldsymbol{V}^{l}\right)=\mathrm{SoftMax}\left(\dfrac{\boldsymbol{Q}^{l}\boldsymbol{K}^{l\mathrm{T}}}{\sqrt{d}}+B\right)\boldsymbol{V}^{l}\end{cases}\tag{2}$$

式中：$\boldsymbol{Q}^{l},\boldsymbol{K}^{l},\boldsymbol{V}^{l}\in\mathbb{R}^{M^2\times d}$ 分别表示的查询（Query）、键值（Key）和值（Value）矩阵；$W_{\boldsymbol{Q}}^{l},W_{\boldsymbol{K}}^{l},W_{\boldsymbol{V}}^{l}$ 表示线性投影层的权重；d 是查询和键值矩阵的维度；M^2 表示每个 2D 窗口的块数量；$B\in\mathbb{R}^{M^2\times M^2}$ 表示每个头的相对位置偏置。

1.2 随机块移位操作

由于烟雾的形状、密度和运动方式可能会因环境条件等因素而有所不同，所以我们认为使用随机块移位模式可增加模型对不同烟雾形态和运动方式的适应能力，提高模型的鲁棒性。本文将时序块移位操作[18]中的固定模式（图 2(a)）改变为随机块移位模式，模型可通过学习不同的时空关系来捕捉烟雾的特征，从而更好地进行烟雾识别，

这将在 2.3 节进行详细的消融实验进行验证。

−4	1	2
−1	0	3
−2	−3	4

(a) pattern-a

$-p4$	$p1$	$p2$
$-p1$	0	$p3$
$-p2$	$-p3$	$p4$

(b) pattern-b

$p8$	$p1$	$p2$
$p7$	0	$p3$
$p6$	$p5$	$p4$

(c) pattern-c

图 2　不同的块移位模式

Fig. 2　Different patch shift patterns

本文设计了简单的随机块移位模式, 如图 2(b)、图 2(c) 所示。其中“0”表示当前帧的块, “$-p$”和“$+p$”分别表示来自当前帧前 p 帧或后 p 帧的块, pattern-b 中的 $p_i \in \{+4,+3,+2,+1\}$, pattern-c 中的 $p_i \in \{\pm4,\pm3,\pm2,\pm1\}$。pattern-b 与 pattern-c 的区别是其采用了均匀的空间分布和对称的帧采样策略, 对应的实验结果将在 2.3 节进行分析。给定一个 2D 窗口 $Z \in \mathbb{R}^{M\times M\times C}$, 窗口中一共有 $M\times M$ 个块, 每个块的维度是 C。随机块移位操作就是将移位模式以滑动窗口的方式重复覆盖至整个 2D 窗口, 随后根据移位大小 p_i 从不同帧中移动相应的块至当前帧, 从而将 2D 的空间自注意力转变为稀疏的时空自注意力。在经过自注意力的计算之后, 再将来自不同帧的块移动回原来位置。

随机块移位以一个零参数、低成本的方式降低了 Transformer 在时序建模中的运算量, 假设每个 2D 窗口中含有 $M\times M$ 个块, 那么全局自注意力、窗口自注意力和具有随机块位移的窗口自注意力的运算量分别如下:

$$\Omega(\text{MSA}) = 4THWC^2 + 2(THW)^2C \tag{3}$$

$$\Omega(\text{WMSA}) = 4THWC^2 + 2PM^2THWC \tag{4}$$

$$\Omega(\text{RPS WMSA}) = 4THWC^2 + 2M^2THWC \tag{5}$$

式中 Video Swin Transformer 中将特征图划分为 $P\times M\times M$ 的 3D 窗口, 那么随机块移位直接将 3D 的自注意力计算转换为 2D 的自注意力。

1.3　可变形注意力

本文在模型中还引入了可变形注意力机制, 使得网络可以自适应地调整注意力的采样位置, 从而能够更好地适应不同形状和尺寸的烟雾目标, 提高烟雾识别的准确性, 减少漏检和误检的情况。图 3 表示可变形注意力与 Swin Transformer 中窗口自注意力的区别, 窗口自注意力需要在每个窗口对所有向量进行自注意力的计算, 而可变形注意力使得自注意力计算仅在偏移后的参考点与查询向量之间进行, 增强了自注意力模块的灵活性, 从而捕捉到更丰富的特征。此外, 在烟雾视频中, 有些区域可能包含了更重要的信息, 例如烟雾的核心区域或者烟雾与其他物体的交界处。可变形注意力机制可通过局部采样的方式, 增强对这些关键区域的关注, 提高对烟雾目标的识别能力。

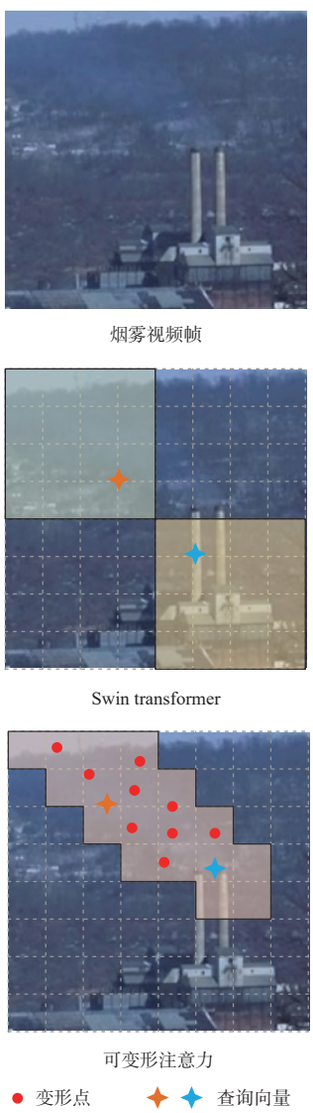


图 3　可变形注意力示意图

Fig. 3　Schematic diagram of deformable attention

如图 4 所示, 首先对输入的特征图 $X \in \mathbb{R}^{H\times W\times 3}$ 做投影得到查询向量 $\boldsymbol{Q}$, 同时生成大小为 $\frac{H}{r}\times\frac{W}{r}\times 2$ 的参考点 p, 参考点的值是线性间隔的二维坐标 $\left\{(0,0),\cdots,\left(\frac{H}{r}-1,\frac{W}{r}-1\right)\right\}$, 然后将它们归一化到 $[-1,1]$。随后将查询向量 $\boldsymbol{Q}$ 送入 Offset 网络学习得到每个参考点的偏移量 Δp, 为了防止偏移量过

大，可以通过参数s来控制（例如$\Delta p \leftarrow s\tanh(\Delta p)$）。将得到的偏移量$\Delta p$与参考点相加得到变形后的参考点，在变形点的位置对特征图进行双线性差值得到$\tilde{X}$，然后投影得到键值向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$。最后对$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$作多头自注意力得到输出，可以用式（6）～式（9）表达：

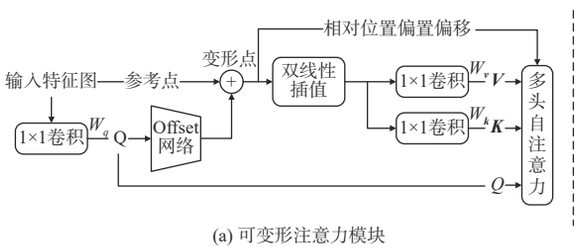
$$\boldsymbol{Q}=XW_Q,\boldsymbol{K}=\tilde{X}W_K,\boldsymbol{V}=\tilde{X}W_V \tag{6}$$

$$\Delta p=\mathrm{Offset}(\boldsymbol{Q}),\tilde{X}=\phi(X;p+\Delta p) \tag{7}$$

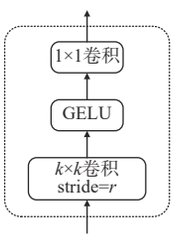
$$z^{(m)}=\mathrm{SoftMax}\left(\frac{Q^{(m)}\boldsymbol{K}^{(m)^{\mathrm{T}}}}{\sqrt{d}}+\phi\left(\hat{B};R\right)\right)V^{(m)},m=1,2,\cdots,h \tag{8}$$

$$Z=\mathrm{Concat}\left(z^{(1)},z^{(2)},\cdots,z^{(h)}\right)W_o \tag{9}$$

式中：$\phi(\cdot)$表示双线性差值；W_o为线性投射层；h为多头自注意力的头部数量；$\phi\left(\hat{B};R\right)$表示可变形的相对位置偏置。



(a) 可变形注意力模块



(b) Offset网络

图 4 可变形注意力

Fig. 4 Deformable attention

2 实验结果与分析

2.1 实验配置

本文的网络结构均采用 Pytorch 深度学习框架实现，编程语言为 Python3.8，使用一张显存为 24 GB 的 RTX3090 显卡进行训练和测试。

实验中所使用的数据集是 RISE（recognizing industrial smoke emissions）数据集，由 HSU Y C 等人[4]开源，是第一个用于识别工业烟雾排放的大规模视频数据集，既包括了从烟囱中排放的烟雾，也包括了工业设备中逃逸的烟雾。该数据集包含了 19 个不同视图的 12 567 个视频片段，每个片段 36 帧（相当于现实世界的 6 min，相机每 5 s～10 s 拍摄一张照片），这些片段监控了 3 个工业设施，共 30 天跨越 2 年，包含 4 个季节，每一个视频被标记为“有烟雾”或者“无烟雾”。该数据集的一大特点是包含了大量的高透明度工业烟雾，这使得基于图像的模型在单帧图像下很难识别出烟雾的动态特性，不具备时序建模的能力，因此只能依靠基于视频的模型来对帧与帧之间的信息进行时空建模，具体的消融实验在 2.3 节进行验证。与其他烟雾数据集不同的是，该数据集涵盖了各种复杂的天气状况，包括雨天、雪天、雷雨等，也掺杂了各种类烟物体，明确区分了蒸汽与烟雾，这对视频烟雾的识别有很强的挑战性。如图 5 所示的数据集概览，直观展示了数据集中一些典型的视频帧，主要包括 (a) 透明度不同的烟雾视频、(b) 大量蒸汽干扰的无烟雾视频、(c) 蒸汽与烟雾混合的视频，其中工业烟雾使用红色箭头标注，蒸汽使用绿色箭头标注。RISE 以 6 种不同方法划分了训练集、验证集和测试集，分别为$S_0,S_1,\cdots,S_5$。其中S_3划分是基于时间，将拍摄视频的前 18 天划分为训练集，之后的 2 天和 10 天用于验证集和测试集。其余 5 种划分方式是基于不同摄像头视角的，并保证每个视角在测试集中至少出现一次，以验证模型在不同视图之间是否稳健。

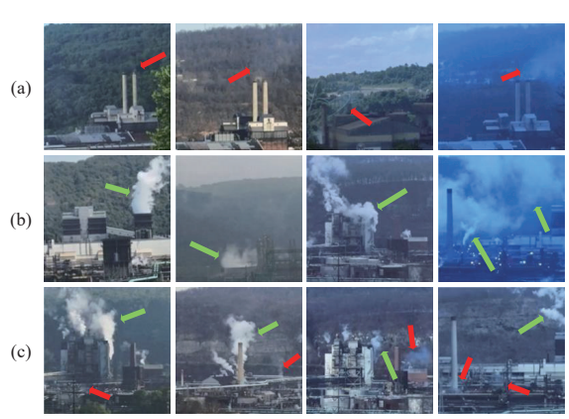


图 5 数据集概览

Fig. 5 Overview of datasets

在训练过程中，应用了标准的数据增强，包括水平翻转、随机调整大小和裁剪。模型采用了二元交叉熵损失函数（binary cross entropy loss, BCE-Loss）和权重衰减自适应矩估计（weight decay adaptive moment estimation, AdamW）进行优化。学习率设置为 0.001，使用余弦退火调整学习率，批量大小为 4，总共训练 40 个轮次。二元交叉熵损失函数

表示为

$$f_{\mathrm{BCELoss}}=-\frac{1}{n}\sum_{i=1}^{n}[y_i\cdot\ln p(y_i)+(1-y_i)\cdot\ln(1-p(y_i))] \tag{10}$$

式中: y_i为二元标签, 值为 0 或 1; $p(y_i)$表示网络预测标签为y_i的输出值。

本文实验采用的主要评价指标是F_1分数($F_{1\mathrm{score}}$), 同时还给出了精确度(P_{r})、召回率(R_{e})、和准确率(A_{cc}), 具体的计算分别如下:

$$P_{\mathrm{r}}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{11}$$

$$R_{\mathrm{e}}=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{12}$$

$$F_{1\mathrm{score}}=\frac{2\times P_{\mathrm{r}}\times R_{\mathrm{e}}}{P_{\mathrm{r}}+R_{\mathrm{e}}} \tag{13}$$

$$A_{\mathrm{cc}}=\frac{T_{\mathrm{P}}+T_{\mathrm{N}}}{T_{\mathrm{P}}+T_{\mathrm{N}}+F_{\mathrm{P}}+F_{\mathrm{N}}} \tag{14}$$

式中:T_{P}表示被正确识别为“有烟雾”; F_{P}表示被错误识别为“有烟雾”; F_{N}表示被错误识别为“无烟雾”; T_{N}表示被正确识别为“无烟雾”。

2.2　实验结果

我们与其他一些主流模型进行了对比, 表 1 展示了不同方法在 RISE 测试集上的F_1分数。可以看到, 本文设计的带有随机块移位和可变形注意力的 Swin Transformer 在S_0、S_2和S_4子集上分别取得了 0.85、0.86 和 0.84 的最佳F_1分数, 这也表明本文所提模型在不同摄像机视角下具有鲁棒性。

表 2 展示了不同模型的参数量(Parameters)、每秒计算的浮点数(FLOPs)和每秒计算帧数(FPS)。本文方法虽然在 FLOPs 上略有增加, 但是保证了较低的参数量以及 31.78 帧/s 的实时处理速度。

为了直观展示本文使用的模型效果, 使用梯度加权类激活映射 Grad-CAM 来可视化网络在输入视频帧中所关注的区域。选取数据集中具有代表性的两类视频帧数据来进行可视化, 分别是具有高透明度的工业烟雾和具有严重蒸汽干扰的工业烟雾, 分别对应图 6 的 (a) 行和 (b) 行的原视频帧, 红色虚线所圈出的是待识别的工业烟雾。如图 6 所示, 2D Swin Transformer 由于不具备时序建模能力, 从而无法正确关注到视频帧中的动态烟雾区域; I3D 模型虽然能正确识别到烟雾, 但是关注的区域相较于本文方法范围更大; 本文使用的模型能够聚焦于视频中的烟雾区域, 并与蒸汽等类烟物体区分开。

表 1　不同方法在 RISE 测试集上比较

Table 1　Comparison of different methods on RISE test set

方法	F_1分数					
	S_0	S_1	S_2	S_3	S_4	S_5
Flow-SVM	0.42	0.59	0.47	0.63	0.52	0.47
Flow-I3D	0.55	0.58	0.51	0.68	0.65	0.50
SVM	0.57	0.70	0.67	0.67	0.57	0.53
I3D	0.80	0.84	0.82	0.87	0.82	0.75
I3D-ND	0.76	0.79	0.81	0.86	0.76	0.68
I3D-FP	0.76	0.81	0.82	0.87	0.81	0.71
I3D-TSM	0.81	0.84	0.82	0.87	0.80	0.74
I3D-LSTM	0.80	0.84	0.82	0.85	0.83	0.74
I3D-TC	0.81	0.84	0.84	0.87	0.81	0.77
CNN-NonFFM[12]	0.83	0.82	0.84	0.85	0.78	0.83
EFFNet[12]	0.84	0.83	0.86	0.86	0.80	0.83
AFSNet[13]	0.85	0.86	0.82	0.91	0.81	0.80
本文方法	**0.85**	0.85	**0.86**	0.88	**0.84**	0.79

表 2　不同方法的性能比较

Table 2　Performance comparison of different methods

方法	Parameters/M	FLOPs/G	FPS
I3D	12.3	62.7	32.71
I3D-TSM	12.3	62.7	31.40
I3D-LSTM	38.0	62.9	32.25
I3D-TC	12.3	62.7	32.88
EFFNet[12]	27.2	34.6	42.57
AFSNet[13]	30.8	40.6	34.87
本文方法	24.2	68.4	31.78

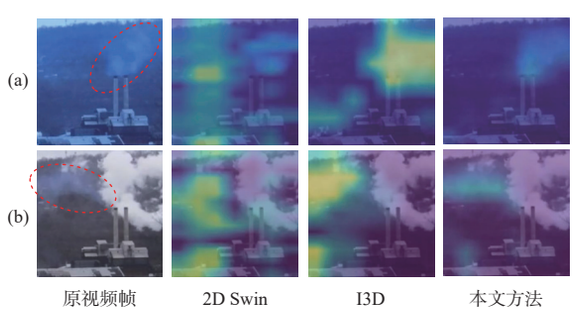


图 6　Grad-CAM 可视化

Fig. 6　Grad-CAM visualization

2.3　消融实验

本文主要研究工作是将随机块移位和可变形注意力引入 Swin Transformer, 来对视频中的烟雾进行高效的时序建模。为了验证其有效性, 在 RISE 数据集的S_0划分子集上设计了详细的消融实验。表 3 展示了带有随机块移位和可变形注意力模型的实验结果。

表 3 RPS 和 DA 消融实验

Table 3 Ablation experiments of RPS and DA

模型	F_1 分数	A_{cc}	P_r	R_e
Swin	0.580 2	0.704 9	0.615 7	0.548 6
Swin+RPS	0.846 5	0.886 9	0.853 8	**0.839 4**
Swin+RPS+DA	**0.850 8**	**0.892 5**	**0.879 2**	0.824 2

表 3 中, Swin 表示最原始的 2D Swin Transformer, 其不具备时间维度上的建模能力, 因此在评价指标上效果很差。在加入了 RPS 之后, Swin 中的空间自注意力会变为稀疏的时空自注意力, 在 F_1 分数和烟雾识别的准确率上都有了大幅度提升。随后, 在 Swin 的最后一层编码器中加入可变形注意力机制, 模型的性能在 F_1 分数、准确率和精确度上又有了小幅度提升, 这是因为 DA 可以根据帧间的不同特征动态调整参考点的偏移量, 这给网络带来了更灵活的感受野, 让网络能够更好地捕捉烟雾的变化特征和运动轨迹, 从而提高识别的精确度。DA 令网络更关注输入特征局部区域的同时, 对全局或更广泛的上下文特征的关联性产生影响, 导致了召回率的下降。从表 3 还可以看到, 针对 Swin 模型, RPS 的加入在召回率的提升上较为显著, 而 DA 的加入更好地实现了精确度与召回率之间的平衡, 从而使得最终的 F_1 分数达到最优。

将 RPS 加入网络中不同层级的编码器进行实验。从表 4 中可以看到, 单加入网络的第 1 层编码器, F_1 分数就比没有 RPS 的网络有了较大提升。通过在网络不同层级的编码器中加入更多的 RPS, 模型的性能也在逐层提高。当 RPS 加入网络的最后一层编码器之后, 模型的性能并没有明显提高, 这是因为 RPS 给网络带来的最大贡献是使得网络能够使用 2D 的自注意力来对视频烟雾进行时序建模, 而网络第 3 层的特征图已经包含了足够丰富的时空信息, 这才导致了 RPS 加入最后一层之后提升不显著。

表 4 RPS 在 Swin Transformer 不同层的消融实验

Table 4 Ablation experiments of RPS in different layers of Swin Transformer

层				F_1 分数	A_{cc}	P_r	R_e
1	2	3	4				
√				0.808 9	0.872 7	**0.915 6**	0.724 4
√	√			0.829 9	0.880 9	0.884 3	0.781 9
√	√	√		0.845 8	**0.888 1**	0.866 9	0.825 9
√	√	√	√	**0.846 5**	0.886 9	0.853 8	**0.839 4**

表 5 展示了随机块移位与 TPS[18] 中的固定移位模式的区别。其中 pattern-a 是 TPS 中 3×3 大小的固定移位模式, 如图 2(a) 所示。pattern-b 则沿用了 pattern-a 中空间均匀分布的规律, 保证了对来自同一帧的块均匀采样。但 pattern-c 能够带来更随机的时间感受野, 而非 pattern-b 对称的时间采样策略。我们认为这种随机移位模式可以增加模型对于烟雾特征提取的泛化性, 有助于提高模型对于烟雾的识别准确性。

表 5 RPS 的不同模式

Table 5 Different patterns of RPS

模式	F_1 分数	A_{cc}	P_r	R_e
pattern-a	0.838 7	0.883 1	0.861 1	0.817 4
pattern-b	0.834 7	0.883 1	**0.880** 0	0.793 7
pattern-c	**0.850 8**	**0.892 5**	0.879 2	**0.824 2**

在表 6 中, 同样将可变形注意力加在网络中的不同阶段, DAT[16] 中效果最好的是加入网络中的最后两层编码器。但本文实验发现, 若加入在最后两层编码器, 会带来性能的下降, 但是若在最后一层引入可变形注意力, 能够在 F_1 分数带来小幅度的提升。分析其原因在于网络的前几个阶段还在学习局部的时空特征, 可变形注意力无法通过局部采样的方式, 增强对关键区域的关注。

表 6 可变形注意力在 Transformer+RPS 不同层的消融实验

Table 6 Ablation experiments of DA in different layers of Swin Transformer+RPS

层		F_1 分数	A_{cc}	P_r	R_e
3	4				
	√	**0.850 8**	**0.892 5**	**0.879 2**	**0.824 2**
√	√	0.809 2	0.866 4	0.863 0	0.761 6

本文以 Swin Transformer 模型框架为基准, 综合以上消融实验, 在 Swin Transformer 的 4 层编码器中分别应用 RPS 模块, 并在最后一层中引入可变形注意力机制, 同时在 RPS 模块中使用 pattern-c 可以使得网络性能达到最优, 与图 1 中所展示的网络框架一致。

3 结论

本文提出了将随机块移位和可变形注意力引入传统的 2D Swin Transformer, 通过将空间自注意力转变为稀疏的时空自注意力来对视频烟雾进行时序建模, 同时凭借可变形注意力对关键区域的

关注来提升视频烟雾识别的准确性。在 RISE 数据集上的实验结果表明，本文提出的方法能够在 3 个子集分别达到 0.85、0.86 和 0.84 的 F_1 分数，相较于其他方法有明显提升。在后续的研究工作中，可以设计更鲁棒的端到端的模型结构，在训练过程中同时考虑精确度和召回率的优化目标，使网络更综合地学习特征表示，从而达到更好的综合性能。

参考文献：

[1] POPE C, DOCKERY D. Health effects of fine particulate air pollution: lines that connect[J]. Journal of the Air & Waste Management Association, 2006, 56(6): 707-708.

[2] 史劲亭, 袁非牛, 夏雪. 视频烟雾检测研究进展[J]. 中国图象图形学报, 2018, 23(3): 303-322.
SHI Jinting, YUAN Feiniu, XIA Xue. Video smoke detection: a literature survey[J]. Journal of Image and Graphics, 2018, 23(3): 303-322.

[3] MIRANDA G, LISBOA A, VIEIRA D, et al. Color feature selection for smoke detection in videos[C]//2014 12th IEEE International Conference on Industrial Informatics (INDIN). New York: IEEE, 2014: 31-36.

[4] HSU Y C, HUANG T H, HU T Y, et al. Project RISE: recognizing industrial smoke emissions[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(17): 14813-14821.

[5] YIN Z, WAN B, YUAN F, et al. A deep normalization and convolutional neural network for image smoke detection[J]. IEEE Access, 2017, 5: 18429-18438.

[6] HE L, GONG X, ZHANG S, et al. Efficient attention based deep fusion CNN for smoke detection in fog environment[J]. Neurocomputing, 2021, 434: 224-238.

[7] LIU Y, QIN W, LIU K, et al. A dual convolution network using dark channel prior for image smoke classification[J]. IEEE Access, 2019, 7: 60697-60706.

[8] ZHOU Y, WANG J, HAN T, et al. Fire smoke detection based on vision transformer[C]//2022 4th International Conference on Natural Language Processing (ICNLP). New York: IEEE, 2022: 39-43.

[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International Conference on Learning Representations (ICLR). [S. l.]: OALiB, 2015: 1-14.

[10] TAO H, LU M, HU Z, et al. Attention-aggregated attribute-aware network with redundancy reduction convolution for video-based industrial smoke emission recognition[J]. IEEE Transactions on Industrial Informatics, 2022, 18(11): 7653-7664.

[11] LIN G, ZHANG Y, XU G, et al. Smoke detection on video sequences using 3D convolutional neural networks[J]. Fire Technology, 2019, 55: 1827-1847.

[12] CAO Y, TANG Q, WU X, et al. EFFNet: enhanced feature foreground network for video smoke source prediction and detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4): 1820-1833.

[13] TAO H, DUAN Q. An adaptive frame selection network with enhanced dilated convolution for video smoke recognition[J]. Expert Systems with Applications, 2023, 215: 119371.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. [S. l.]: [s. n.], 2017: 5998-6008.

[15] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 10012-10022.

[16] XIA Z, PAN X, SONG S, et al. Vision transformer with deformable attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 4794-4803.

[17] LIU Z, NING J, CAO Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 3202-3211.

[18] XIANG W, LI C, WANG B, et al. Spatiotemporal self-attention modeling with temporal patch shift for action recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 627-644.