arXiv:2603.04882v1 [cs.CV] 5 Mar 2026

DeformTrace：用于时序伪造定位的中继令牌可变形状态空间模型

Xiaodong Zhu、Suting Wang、Yuanming Zheng、Junqi Yang、Yangxu Liao、Yuhong Yang*、Weiping Tu、Zhongyuan Wang

多媒体软件国家工程研究中心、计算机科学学院，武汉大学，中国；xiaodongzhu@whu.edu.cn、yangyuhong@whu.edu.cn、wzy hope@163.com

摘要

时序伪造定位（TFL）旨在精确识别视频和音频中的篡改分段，为安全与取证提供强有力的可解释性。尽管近期的状态空间模型（SSM）在精确时序推理方面展现出良好前景，但边界模糊、伪造稀疏以及长程建模能力有限等问题，阻碍了其在 TFL 中的应用。我们提出 DeformTrace，通过可变形动态和中继机制增强 SSM，以应对这些挑战。具体而言，可变形自状态空间模型（DS-SSM）为 SSM 引入动态感受野，以实现精确的时序定位。为进一步增强其时序推理能力并缓解长程衰减，我们在 DS-SSM 中集成了中继令牌机制。此外，可变形交叉状态空间模型（DC-SSM）将全局状态空间划分为查询特定子空间，从而减少非伪造信息的累积，并提高对稀疏伪造的敏感度。这些组件被集成到一种混合架构中，结合了 Transformer 的全局建模能力与 SSM 的效率。大量实验表明，DeformTrace 以更少的参数、更快的推理速度和更强的鲁棒性，实现了最先进的性能。

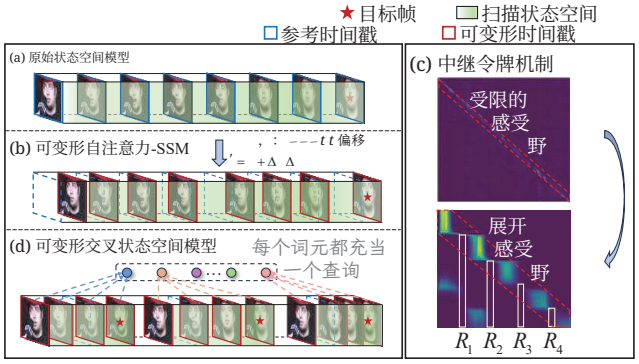


图 1：我们主要贡献的概览（以视频序列为例进行说明）。(a) 原始状态空间模型；(b) 具有可学习时间偏移量的可变形自状态空间模型，可实现灵活的局部采样；(c) 隐式注意力可视化，其中中继令牌扩展感受野并维持长程依赖；(d) 可变形交叉状态空间模型允许每个查询词元将全局状态空间划分为多个子空间，从而实现跨序列交互。

引言

生成式人工智能的快速发展极大地推动了多媒体创作，但由于逼真伪造内容易于生成，也引发了安全方面的担忧。现有大多数研究（Feng、Chen 和 Owens，2023；Oorloff 等人，2024；Smeu 等人，2025）主要关注二元伪造检测，而时序伪造定位（TFL）则通过识别被篡改的片段，提供了更细致的可解释性。

早期的时序伪造定位方法，例如 BA-TFD（Cai 等人，2022）和 BA-TFD+（Cai 等人（2023））， 依赖卷积神经网络或多尺度 Transformer进行帧级分类，但存在精确度低、推理速度慢的问题。后来的模型，如 UMMAFormer（Zhang 等（2023））和 DiMoDif（Koutlis 和 Papadopoulos，2024），借助预训练模型提升了性能

提取器与特征金字塔，但仍计算开销沉重。因此，开发一种紧凑而有效的TFL架构势在必行。TFL在概念上与时序动作检测（TAD）相关，因为二者都需要精确的边界在长序列中的定位。TAD利用状态空间模型（SSMs）和基于查询的设计所取得的最新进展，为模型设计提供了有价值的启示。

近期，状态空间模型（SSMs）取得了显著进展，尤其是 Mamba （ Gu and Dao 2024 ； Dao and Gu 2024 ），在复杂度更低、推理速度更快的同时展现出强劲性能。MambaIRv2（Guo et al. 2025）和 ActionMamba（ Chen et al. 2024）等模型在分割和时间动作检测等任务中的表现优于基线方法，展示了状态空间模型在密集、时间感知预测中的潜力。然而，由于以下三大挑战，基于状态空间模型的架构尚未应用于 TFL：**边界模糊性**：伪造边界通常并不清晰，不像 TAD 中的边界那样定义明确。标准状态空间模型采用固定状态更新，会造成时间平滑，从而降低定位精度；**稀疏伪造**：大多数帧都未被伪造，因此状态空间模型的递归更新会受到非伪造模式主导，削弱对稀疏伪造的敏感性；**长序列建模能力有限**：尽管状态空间模型在处理长序列时效率很高，但会出现信息随距离衰减的现象（Ye et al. 2025），限制其

*通讯作者。版权所有 © 2026，人工智能促进协会 (www.aaai.org)。保留所有权利。

捕获长程上下文的能力。

为解决边界模糊性，我们提出可变形自状态空间模型（DS-SSM），首次将可变形动态感受野机制引入状态空间模型。不同于图像领域的可变形 Mamba 变体（Liu 等人（2025）；Hu 等人（2024）），我们的设计利用视频和音频固有的时间连续性，因此省去了图像块拆分和词元排序等操作。相反，DS-SSM 在每个时间步预测偏移量，以动态采样输入特征，在保留状态空间模型低复杂度的同时，显著降低计算开销。图 1(b) 显示，与原始状态空间模型（见图 1(a)）相比，DS-SSM 能更好地捕捉局部固定窗口之外的语义相关上下文，并增强对模糊时间边界的鲁棒性。

与此同时，受无线通信中中继节点的启发，我们在 DS-SSM 中引入中继令牌机制，以维持长程信息流。如图 1(c) 所示，该图比较了原始状态空间模型与我们加入中继令牌的变体的隐式注意力（可视化方式参照 Ali、Zimerman 和 Wolf（2025）），这些按周期插入的可学习令牌能有效扩大感受野，并缓解长程衰减问题。

我们进一步提出可变形交叉状态空间模型（DC-SSM），将跨序列交互引入可变形状态空间建模，以应对稀疏伪造。如图 1(d) 所示，每个辅助令牌代表一种潜在伪造，并作为查询，从主流序列中检索与伪造相关的信息。该机制将全局状态空间划分为查询特定子空间，减少非伪造内容的累积，并提高对稀疏伪造的敏感度。与交叉注意力类似，DC-SSM 支持显式的令牌到序列交互，以实现定向检索。最后，这些组件被集成到一种混合 TFL 架构中，将 Transformer 注意力与高效的状态更新相结合。

总而言之，本工作的主要贡献如下：（1）我们提出可变形自状态空间模型（DS-SSM），首次将动态感受野引入时序状态空间模型，从而改善对模糊边界的定位。（2）我们引入中继令牌机制，明确缓解状态空间模型的长程衰减；这是以往状态空间模型的一项关键局限。（3）我们提出可变形交叉状态空间模型（DC-SSM），首次将跨序列交互融入状态空间建模，从而增强对稀疏伪造的敏感度。（4）我们将这些组件整合为统一的混合 TFL 架构，结合 Transformer 与状态空间模型的优势。在标准基准数据集上的大量实验表明，该方法以更小的模型规模和更快的推理速度实现了最先进的性能。

相关工作

状态空间模型应用

状态空间模型（SSMs）为序列建模提供了一种高效的方法。Mamba（Gu 和 Dao，2024年；Dao 和 Gu，2024年）在此基础上引入选择机制，实现了具有线性复杂度的长序列建模。Mamba 推动了越来越多面向视觉任务的变体的发展，这些变体着重于可扩展性、效率和局部-全局建模（Zhu 等人，2024年；Liu 等人，2024年b；Pei、Huang 和 Xu，2025年；Patro 和 Agneeswaran，2024年；Hatamizadeh 和 Kautz，2025年）。与此同时，可变形变体（Liu 等人，2025年；Hu 等人（2024））采用可变形扫描来处理不规则模式，但仅限于单向序列建模。此外，Mamba 及其变体都存在长程衰减问题。LongMamba（Ye 等人，2025年）尝试通过通道分类和词元过滤来解决这一问题，但其依赖固定阈值，限制了适应性。

时间动作检测

时间动作检测（TAD）旨在检测并定位未剪辑视频中的动作。早期的基于锚点的方法（Bai 等，2020年；Lin 等，2018年、2019年；Long 等，2019年）由于固定锚点，难以应对不同的动作时长。无锚点方法（Cheng 和 Bertasius，2022；Lin 等，2021；Zhang、Wu 和 Li，2022）通过基于中心或非对称设计提高了灵活性。受 DETR（Carion 等，2020）启发，TadTR（Liu 等，2022）和 TE-TAD（Kim 等，2024年）等基于查询的方法将时间动作定位视为集合预测，从而移除锚点和后处理（Tan 等，2021）。近年来，基于Mamba的架构（Chen 等，2024年；Liu 等，2024年a；Wang 等，2025年）被引入 TAD，增强了时序建模能力。

时序伪造定位

时序伪造定位（TFL）旨在识别视频中的篡改片段，需要细粒度时间建模。尽管大多数研究聚焦于视频级检测（Zhou 与 Lim，2021；Yang 等人，2023；Yu 等人，2023），TFL 仍未得到充分探索。近期的数据集，如 Lav-DF（Cai 等人（2023））和 AV-Deepfake1M（Cai 等人（2024）），已开始推动这一方向的进展。BA-TFD（Cai 等（2022））引入了用于帧级分类的对比损失，随后 BA-TFD+（Cai 等人（2023））采用了多尺度 Transformer。UMMAFormer（Zhang 等人，2023）、DiMoDif（Koutlis 与 Papadopoulos，2024年）、MFMS（Zhang 等人，2024年）和 CLFormer（Cheng 等人，2025年）通过注意力和特征金字塔提高了准确率。Vigo（P´erez-Vieites 等人，2024年）采用基于IoU的融合策略，而 TransHFC（Huang 等人，2025年）则融合了超图学习与 Transformer。Zhu 等人（Zhu 等人，2025年）提出了一种基于查询的音视频框架，以实现高效的端到端时间定位。

所提方法

总体架构

给定一个未剪辑说话人脸视频数据集，每个样本 $x = \{x^v, x^a\}$ 包含时间对齐的视频和音频流。然而，x 的部分内容可能在任一模态中被篡改。DeformTrace的目标是定位伪造片段 $\mathcal{A} = \{ (t_i, d_i) \}_{i=1}^{N_f}$，其中 t_i 和 d_i 分别表示第 i个伪造的中心时间戳和持续时间。此处，N_f 表示伪造的总量。此外，它还预测一个二值真实性标签 y。

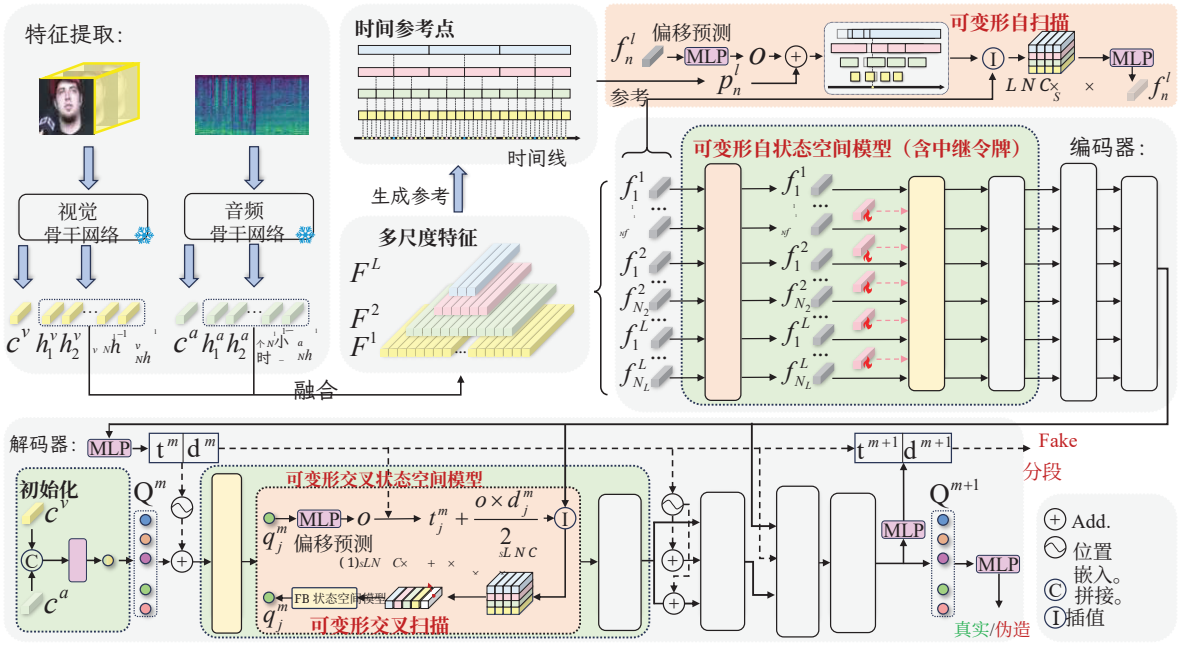


图 2：DeformTrace 整体方案示意图。DeformTrace 基于 TadTR（Liu 等，2022）构建，集成了多尺度音视频特征提取模块、用于时序建模的可变形编码器，以及用于伪造定位和视频级分类的可变形解码器。通过引入可变形自状态空间模型和交叉状态空间模型模块，它将 Mamba 高效的状态更新与 Transformer 的全局建模相结合。增强损失和协作损失所辅助的中继令牌，有助于在自扫描期间保留长程信息依赖关系。

特征提取。 如图2所示，DeformTrace采用基于TadTR（Liu等人，2022）的基于查询的架构。视频 $x^v \in \mathbb{R}^{T\times H\times W\times C_0}$和梅尔频谱图 $x^a \in \mathbb{R}^{T\times N_{\text{mel}}}$ 由冻结的预训练视觉与音频骨干网络处理，以提取分类词元 $\{c^v, c^a\}$ 和特征序列 $\{h_n^v, h_n^a\}_{n=1}^{N_1}$，其中分类词元与特征词元的维度均为 C。此处， T 和 N_1 分别表示输入帧数以及第一尺度提取的词元数量。为保留用于细粒度定位的时间分辨率，我们设定 $T = N_1$。这些序列经拼接并线性投影，形成融合特征 F^1，再由（$L-1$）2×下采样层进一步处理，得到特征$\mathcal{F} = \{F^l\}_{l=1}^{L}$，从而捕获多尺度时序上下文。

编码器。 编码器接收展平后的特征$\{f_n^l\}_{n=1}^{N_l}$，其中 n 表示各尺度的特征索引， $N_l = N_1/2^{l-1}$， $l = 1, 2, \cdots, L$，随后编码器输出上下文增强序列 X_E。如图 2 所示，它由可变形自状态空间模型、可变形自注意力（DSA）（Liu 等，2022）和前馈网络（FFN）组成。DSA 会自适应地关注参考点周围稀疏的时序位置，而 FFN 则引入非线性并进一步优化特征。

解码器。视觉和音频分类令牌通过多层感知机融合为全局多模态特征，并重复 N_q 次以形成初始查询 $Q^0 = \{q_j^0\}_{j=1}^{N_q}$，从而与真值实现粗略对齐并促进收敛（Kim 等，2024）。初始候选区域 $\mathcal{A}^0 = \{(t_j^0, d_j^0)\}_{j=1}^{N_q}$表示分段的中心和持续时间，由融合特征经多层感知机生成。解码器包含M 层，每层由可变形交叉状态空间模型、多头自注意力（MHSA）、可变形交叉注意力（DCA）（Liu 等，2022）和前馈网络组成。MHSA 使用基于锚点的位置嵌入对查询之间的关系进行建模，而 DCA 则使每个查询能够关注相关的音频-视觉上下文。DCA 的输出经前馈网络处理以更新查询，随后通过多层感知机预测偏移量 Δt， Δd，用于候选框精细化。最终候选区域 $\mathcal{A}^M$ 作为预测的伪造片段，最终查询Q^M 则经多层感知机分类器处理，生成视频级伪造分数 $\hat{y}$。

可变形自状态空间模型

可变形自状态空间模型（DS-SSM）模块由三个部分组成：可变形自扫描、前向-后向状态空间模型（FB 状态空间模型）（Wang等人2023）以及前馈网络（FFN）。可变形自扫描使每个词元能够灵活地采样局部窗口之外的语义相关位置。FB 状态空间模型在采样序列上执行双向状态空间传播，从过去和未来捕获上下文信息。

对于多尺度特征 $\mathcal{F}$第 l层中的第 n个特征（记为 f_n^l），我们计算归一化参考位置p_n^l， 以与视频的时间轴对齐：

$$p_n^l = \frac{\omega \cdot 2^{l-1} \cdot (n + 0.5)}{\text{fps} \cdot d}. \tag{1}$$

其中，fps表示每秒帧数（默认为25），ω 为步长，表示提取特征时每次步进的帧数，d 为以秒计的视频时长。上述分子以帧索引估计 f_n^l 的时间中心，除以fps将其转换为秒，再除以 d进行归一化，得到位于$(0,1]$内的值。我们预先计算这些时间参考点，以加速采样过程。

给定特征 f_n^l 及其时间参考点 p_n^l，我们通过多层感知机预测偏移矩阵 $o \in \mathbb{R}^{L\times N_s}$，其中 N_s 为每个尺度的偏移数量。对于每个特征，会在多个尺度上预测多个偏移，从而增强采样的灵活性与多样性。p_n^l 被重复并与 o 逐元素相加，以得到可变形点$\hat{}p$。双线性插值函数 $\phi(\cdot,\cdot)$ 从多尺度特征 $\{F^l\}_{l=1}^L$中提取位置$\hat{}p$处的特征，再通过另一个多层感知机进行聚合，生成输出特征$\hat{}$ f_n^l。该过程表述为：

$$\begin{aligned} o &= \mathrm{MLP}_1(f_n^l), \\ \hat{p} &= \mathrm{Repeat}(p_n^l) + o, \\ \hat{f}_n^l &= \mathrm{MLP}_2\left(\phi\left(\hat{p}, \{F^l\}_{l=1}^L\right)\right). \end{aligned} \tag{2}$$

中继令牌机制

长程衰减是状态空间模型的固有局限。词元交互依赖于控制矩阵的幂 $\bar{A}^k$，其中 k 为成对距离（Gu和Dao 2024年；Dao和Gu 2024年）。¯由于 A包含小于1的值，交互会随距离呈指数衰减。例如，在DS-SSM中，语义对齐的词元 f_1^1与 f_1^L相隔 $k \approx 2N_1$（在AV-Deepfake1M和LAV-DF等数据集中 $N_1 > 100$），导致交互可忽略且感受野受限（见图 1(d)）。

为解决这一问题，我们从通信系统中获得启发：中继节点可缓解长距离信号衰减。我们提出中继令牌机制：可学习的全局令牌 $\{r_i\}_{i=1}^{N_r}$ 在状态空间更新之前被均匀插入输入序列。这些与输入无关的令牌将序列划分为 N_r+1 个子空间。在每个子空间内，局部状态高效地将信息传递给中继令牌，再由该令牌将聚合后的消息广播到其他子空间。由此在整个序列上形成稀疏序列到令牌的信息流。图 1(d) 展示了注意力图，说明感受野的扩展以及经由中继令牌实现的信息中继。

可变形交叉状态空间模型

可变形交叉状态空间模型（DC-SSM）模块由三个组件构成。FB-SSM 捕获每个查询周围的时间依赖，而可变形交叉扫描使每个查询能够动态关注来自编码器的语义相关音视特征，从而增强对伪造区域的定位。前馈网络进一步细化所得特征。

在可变形交叉扫描中，对于解码器的第 m 层，每个查询 $q_j^m \in Q^m$ 都使用其锚点提议 $(t_j^m,\ d_j^m)$ 作为参考。一个多层感知机预测多尺度偏移量 $o \in \mathbb{R}^{L\times N_s}$，并据此计算可变形点 $p_j^m = t_j^m + o \times d_j^m/2$。双线性插值函数从编码器输出 X_E 中采样 p_j^m 处的特征，生成形状为 $L \times N_s \times C$ 的特征张量。该张量经展平后与一个可学习空令牌拼接，形成长度为 $(L \times N_s + 1) \times C$ 的序列。随后，将该序列输入前向状态空间更新。由于状态空间模型的聚合特性，只保留最后一个追加的词元作为输出特征 $\hat{}q_j^m$。

损失与优化

在中继令牌机制中，我们引入了两项辅助损失：增强损失和协作损失。前者促使每个中继令牌更好地聚合其邻近序列片段中的信息，而后者促进中继令牌之间的有效协作。

增强损失。对于第 k个中继令牌 r_k，我们计算其邻近非中继子序列（不包括其他中继令牌）的平均表示 f_k^{avg}。增强损失随后定义为：

$$\mathcal{L}_{\text{enh}} = -\frac{1}{N_r}\sum_{k=1}^{N_r}\cos\left(r_k, f_k^{avg}\right), \tag{3}$$

协作损失。为促进多样性并减少中继令牌之间的冗余，我们引入协作损失，该损失最小化不同中继令牌之间的互信息（MI）。这可以通过促使中继令牌的相似度矩阵 G趋近于一个缩放后的单位矩阵来近似：

$$\begin{aligned} \mathcal{L}_{\text{coop}} &= \min\sum_{i<j} I(r_i; r_j), \\ &\approx \|G - \gamma I\|^2 = \sum_{i\neq j} G_{ij}^2 + \sum_{k=1}^{N_r}(G_{kk}-\gamma)^2, \end{aligned} \tag{4}$$

其中 γ 是控制目标自相似度的超参数，我们将其设为1。

任务感知损失。遵循DETR（Carion等人，2020），我们采用标准的匈牙利匹配损失 $\mathcal{L}_{\text{match}}$，其中包括分类损失 $\mathcal{L}_{\text{ce}}$ 和回归损失$\mathcal{L}_{\text{reg}}$。其定义为：

$$\mathcal{L}_{\text{match}}(\mathcal{A}^M, \mathcal{A}) = \sum_{i=1}^{N_f}\sum_{u\in\{\text{ce,reg}\}}\mathcal{L}_u\left(\mathcal{A}^M_{\pi(i)}, \mathcal{A}_i\right), \tag{5}$$

其中，π 表示通过匈牙利匹配得到的最优排列，$\mathcal{A}^M$，$\mathcal{A}$ 分别表示预测分段和真实标签分段。此外，我们对预测标签 $\hat{}y$ 与真实标签 y计算视频级交叉熵损失，记作 $\mathcal{L}_{\text{cls}}$。

总体而言，训练损失由以下四个部分加权组合而成：

$$\mathcal{L} = \mathcal{L}_{\text{match}} + \mathcal{L}_{\text{cls}} + \lambda_1\cdot\mathcal{L}_{\text{enh}} + \lambda_2\cdot\mathcal{L}_{\text{coop}}. \tag{6}$$

其中，λ_1 和 λ_2 是超参数，分别用于控制增强损失和协作损失的贡献。

Type	方法	发表 venue	模态	平均精度均值(%)				平均召回率(%)				
				0.5	0.75	0.95	平均值	100	50	20	10	Avg.
锚点-基于	BMN	ICCV 2019	$\mathcal{V}$	24.0	7.60	0.10	10.6	53.3	41.2	31.6	26.9	38.3
	BMN（I3D）	ICCV 2019	$\mathcal{V}$	10.6	1.70	0.00	4.10	48.5	44.4	37.1	31.6	40.4
锚点-free	AGT	Arxiv'21	$\mathcal{V}$	17.9	9.40	0.10	9.10	43.2	34.2	24.6	16.7	29.7
	ActionFormer	ECCV'22	$\mathcal{V}$	95.3	90.2	23.7	69.7	88.4	89.6	90.3	90.4	89.7
	TriDet	CVPR'23	$\mathcal{V}$	96.3	86.8	23.6	68.9	91.0	90.4	90.0	88.7	90.0
	MDS	MM'20	$\mathcal{AV}$	12.8	1.60	0.00	4.80	37.9	36.7	34.4	32.2	35.3
	AVFusion	Arxiv'21	$\mathcal{AV}$	65.4	23.9	0.10	29.8	63.0	59.3	54.8	52.1	57.3
	BA-TFD	DICTA'22	$\mathcal{AV}$	76.9	38.5	0.30	38.6	66.9	64.1	60.8	58.4	62.5
	BA-TFD+	CVIU'23	$\mathcal{AV}$	96.3	85.0	4.40	61.9	81.6	80.5	79.4	78.8	79.8
	UMMAFormer	MM'23	$\mathcal{AV}$	**98.8**	**95.5**	37.6	**77.3**	92.4	92.5	92.5	92.1	92.3
	ELF-MDC	TCSVT'24	$\mathcal{AV}$	94.9	74.9	1.90	57.2	76.1	74.2	72.2	71.2	73.4
	DiMoDif	Arxiv'24	$\mathcal{AV}$	95.5	87.9	20.6	67.8	**94.2**	**93.7**	92.7	91.4	91.9
查询-基于	TadTR	TIP'22	$\mathcal{V}$	80.2	61.1	5.20	48.8	72.9	72.5	70.56	69.2	71.2
	TE-TAD	CVPR'24	$\mathcal{V}$	85.7	64.9	7.13	52.6	79.2	78.6	78.4	76.2	77.9
	FullFormer（我们的）	–	$\mathcal{AV}$	94.6	[illegible]	29.4	69.9	88.4	[illegible]	86.9	85.5	87.3
	DeformTrace（我们的方法）	–	$\mathcal{AV}$	97.1	[illegible]		75.3	93.3	[illegible]	**92.8**	**92.3**	**92.9**

表 1：LAV-DF（Cai 等人（2023））基准测试集上的时序伪造定位结果。模态表示输入类型：$\mathcal{V}$表示视觉，$\mathcal{A}$ 表示音频，$\mathcal{AV}$ 表示两者兼有。最佳结果以**粗体**标示，次佳结果加下划线。

实验

实验设置

数据集与评估指标。我们在两个音视频深度伪造数据集上进行实验：LAV-DF（Cai 等人（2023））和 AV-Deepfake1M（Cai 等人（2024））。LAV-DF 包含 78K/31K/26K 个训练/验证/测试样本。AV-Deepfake1M 是规模更大、挑战性更高的基准数据集，包含 746K/57K/343K 个样本和更精细的伪造内容，对时序伪造定位模型提出了更高要求。按照标准做法，我们报告交并比阈值 $\{0.5, 0.75, 0.9, 0.95\}$ 下的平均精度均值（mAP），以及不同提议数量 $\{5, 10, 20, 30, 50, 100\}$ 下的平均召回率均值（mAR）。此外，我们使用曲线下面积（AUC）作为指标，评估 AV-Deepfake1M 上的视频级检测性能。我们在单张 RTX 3090 GPU 上比较模型参数量和计算成本，并在完整的 LAV-DF 测试集（视频平均时长：8.6 秒）上评估所有方法，统计总参数量/可训练参数量、浮点运算次数和推理时间。

实现细节。由于 Raven 的视觉和音频编码器（Haliassos 等人（2023））在多种下游任务中表现出色，我们采用它们进行特征提取。具体而言，我们使用在 VoxCeleb2（Chung、Nagrani 和 Zisserman（2018））和 LRS3（Afouras、Chung 和 Zisserman（2018））上预训练的基础模型。视频以 25 帧每秒采样，音频采样率为 16 kHz。视觉帧会调整为 224×224 像素，音频则使用 40 ms 汉明窗和 40 ms 的跳步转换为 80 个频带的梅尔频谱图。每个输入片段时长为 8 秒，编码后得到 $T = N_1 = 200$ 帧和 $C = 256$ 维特征。我们通过核大小为 3 的一维卷积采用 $L = 6$ 个特征层级，并使用 $M = 3$ 个解码器层。在每个可变形自/交叉状态空间模型层中，我们在每个尺度采样 $N = 6$ 个点。查询和中继令牌的数量分别设为 $N_q = 60$ 和 $N_r = 8$。总损失由以下两项加权组合而成两项，权重分别为 $\lambda_1 = 0.5$ 和 $\lambda_2 = 0.2$。我们在 AV-Deepfake1M 上训练 30 轮，在 LAV-DF 上训练 100 轮，批次大小为 32，使用 AdamW（Loshchilov 和 Hutter，2019年）优化器和带 5 轮预热的余弦学习率调度器。学习率设为 2e-4。所有实验均在八块 NVIDIA RTX 3090 GPU 上进行。

比较方法。对于时序伪造定位任务，我们的基线系统包括三种公开可用的音视时序伪造定位方法：BA-TFD（Cai 等（2022））、BA-TFD+（Cai 等人（2023））和 UMMAFormer（Zhang 等（2023））。此外，我们将 DeformTrace 中的所有状态空间模型替换为 Transformer 模块，构建了一个基线系统，称为 FullFormer，以供比较。为进行全面比较，我们还报告了在 LAV-DF 和 AV-Deepfake1M 基准数据集上的结果。

与当前最先进方法的比较

时序伪造定位（TFL）结果。表 1 和表 2 对比了我们的方法与先进的时序伪造定位方法在 LAV-DF 和 AV-Deepfake1M 上的表现。在这两个数据集上，我们的方法在平均精度均值和平均召回率方面均比纯 Transformer 基线高出 7% 以上，凸显了状态空间模型和我们提出的模块的有效性。在 LAV-DF 上，该方法在 mAP@0.95 和 mAP@{20, 10, Avg.} 上取得最佳结果，并在其余指标上排名第二，展现出在不同评估标准下的稳健性。在 AV-Deepfake1M 上，由于数据多样性更高且分段更短，UM-MAFormer 和 BA-TFD+ 等领先模型的性能有所下降。相比之下，我们的方法始终取得先进的结果，在所有指标上均以显著优势超过现有方法。具体而言，DeformTrace 在 mAP@{0.5、0.75、0.9 和 0.95} 上分别比排名第二的模型 DiMoDif 高出 5.1%、2.7%、2.5% 和 4.15%，平均提升 3.6%。在使用 50、30、20、10 和 5 个候选区域时，Deform-

Type	方法	发表 venue	模态	平均精度均值(%)					平均召回率(%)					
				0.5	0.75	0.9	0.95	平均值	50	30	20	10	5	Avg.
锚点-free	Meso4	WIFS'18	$\mathcal{V}$	9.86	6.05	2.22	0.59	4.68	38.9	38.9	38.8	36.5	26.9	36.0
	MesoInception4	WIFS'18	$\mathcal{V}$	8.50	5.16	1.89	0.50	4.01	39.3	39.2	39.0	35.8	24.6	35.6
	EfficientViT	ICIAP'22	$\mathcal{V}$	14.7	2.42	0.13	0.01	4.32	27.1	27.0	26.4	23.9	20.3	24.9
	TriDet+VideoMAEv2	CVPR'23	$\mathcal{V}$	21.7	5.83	0.54	0.06	7.03	20.3	20.2	20.1	19.5	18.2	19.7
	TriDet+InternVideo	CVPR'23	$\mathcal{V}$	29.7	9.02	0.79	0.09	9.89	24.1	24.1	24.0	23.5	22.6	23.6
	ActionFormer+VideoMAEv2	ECCV'22	$\mathcal{V}$	20.2	5.73	0.57	0.07	6.65	20.0	19.9	19.8	19.1	17.8	19.3
	ActionFormer+InternVideo	ECCV'22	$\mathcal{V}$	36.1	12.0	1.23	0.16	12.4	27.1	27.1	27.0	26.6	25.8	26.7
	BA-TFD	DICTA'22	$\mathcal{AV}$	37.4	06.3	0.19	0.02	11.0	45.6	40.4	36.0	30.7	26.8	35.9
	BA-TFD+	CVIU'23	$\mathcal{AV}$	44.4	13.6	0.48	0.03	14.6	48.9	44.5	40.4	34.7	29.9	39.7
	UMMAFormer	MM'23	$\mathcal{AV}$	51.6	28.1	7.65	1.58	22.2	44.1	43.9	43.5	42.1	40.3	42.8
	DiMoDif	Arxiv'24	$\mathcal{AV}$	86.9	76.0	28.7	5.43	49.3	81.6	80.9	80.3	78.8	76.6	79.6
基于查询的方法	TadTR	TIP'22	$\mathcal{V}$	60.5	33.8	5.12	1.20	25.2	57.6	55.8	54.2	51.2	50.0	53.8
	TE-TAD	CVPR'24	$\mathcal{V}$	71.4	54.4	11.3	2.31	34.9	66.8	64.7	64.0	61.4	59.9	63.3
	FullFormer（我们的方法）	–	$\mathcal{AV}$	77.3					70.8					
	DeformTrace（我们的方法）	–	$\mathcal{AV}$	**92.0**					**86.2**					

表 2：AV-Deepfake1M（Cai 等人（2024））基准测试集上的时序伪造定位结果。模态表示输入类型：$\mathcal{V}$ 视觉，$\mathcal{A}$ 音频，$\mathcal{AV}$ 两者兼有。最佳结果以 **粗体**标示，次优结果加下划线。

方法	总量[M]	训练[M]	浮点运算次数[G]	时间[毫秒]
BA-TFD	**5.5**	**5.5**	948.1	605
BA-TFD+	152.9	152.9	218.2	681
UMMAFormer	165.9	49.72	1563.9	857
FullFormer（我们的方法）	116.7	19.2	233.7	126
DeformTrace（我们的方法）	118.3	20.8	**212.4**	**104**

表3： $\mathcal{AV}$TFL 方法的模型规模与计算量比较。总量、训练和时间 分别表示总参数量、可训练参数量和单视频推理时间。

#	DS-SSM	DC-SSM	$\mathcal{L}_{\text{enh}}$	$\mathcal{L}_{\text{coop}}$	mAP	mAR	AUC
原始状态空间模型							
1	✗	✗	✗	✗	41.2	68.7	90.3
可变形状态空间模型							
2	✓	✗	✗	✗	44.4	73.2	95.7
3	✗	✓	✗	✗	47.8	75.7	97.3
4	✓	✓	✗	✗	49.8	78.2	98.1
中继令牌机制							
5	✓	✓	✓	✗	51.4	79.2	98.3
6	✓	✓	✗	✓	51.7	78.9	98.5
7	✓	✓	✓	✓	**52.9**	**81.8**	**99.2**

表4：AV-Deepfake1M上的消融研究结果。

Trace 的表现分别提升了 4.6%、3.3%、1.1% 和 1.4%，平均比 DiMoDif 提升 2.2%。这些结果清楚地表明，即使在具有挑战性的条件下，该模型也能精准定位细粒度的伪造分段，充分展现了其优势。

效率结果。我们比较了音视时序伪造定位方法的模型规模和计算成本，并报告总参数量与可训练参数量，以考虑特征提取器之间的差异。计算成本和推理时间均包括特征提取。如表3所示，我们的模型效率更高、更加轻量：与 UMMAFormer 和 BA-TFD+ 相比，可训练参数量分别减少了 28.92M 和 132.1M，且 GFLOPs 最低，比 UMMAFormer 少 6.4×。此外，其运行速度比 UMMAFormer 快 7.3×，比 BA-TFD 和 BA-TFD+ 快 5.8× 。与我们的基线相比，DeformTrace 的参数量略有增加，但计算量和推理时间显著减少。这种效率优势得益于基于查询的架构（完全端到端，无需后处理）以及状态空间模型的线性复杂度。模型的计算成本主要集中在序列处理模块，其中线性复杂度的状态空间模型中 DS-SSM 与 DC-SSM 的成本比约为 4:1，而轻量级的基于多层感知机的偏移预测所占比重可以忽略不计。总的来说，DeformTrace 在性能与效率之间实现了更好的权衡。

消融研究

DeformTrace 的消融研究。为评估各组件的影响，我们在 AV-Deepfake1M 完整测试集上进行了消融实验，如表4所示。与基线模型（第1行）相比，基线模型将所有可变形状态空间模型替换为原始状态空间模型，并移除了中继损失；DeformTrace（第7行）的平均精度均值提升了 28.4%，平均召回率提升了 19.1%，曲线下面积提升了 9.9%，凸显了所有模块的有效性。具体而言，第2行和第3行表明，DS-SSM 和 DC-SSM 均能提升性能，其中 DC-SSM 带来的增益更大，原因在于：1）缺少中继令牌时，DS-SSM 受到长程衰减的影响更大；2）相较于原始状态空间模型，DC-SSM 增强了查询-特征序列交互，这对候选框精细化至关重要。第4行表明，结合这两个模块可进一步提升性能。比较第4行与第5至第6行可知，每种中继损失都能带来适度的增益，而同时使用两者（第7行）效果最佳，平均精度均值、平均召回率和曲线下面积分别比第4行高出 3.1%、3.6% 和 1.1%。

超参数消融研究。 我们针对中继令牌数量 N_r进行消融研究，使用平均视频时长为9秒的AV-Deepfake1M fulltest集。如图4(a)所示，当 $N_r = 1$或为2时，性能与基线相当或更差

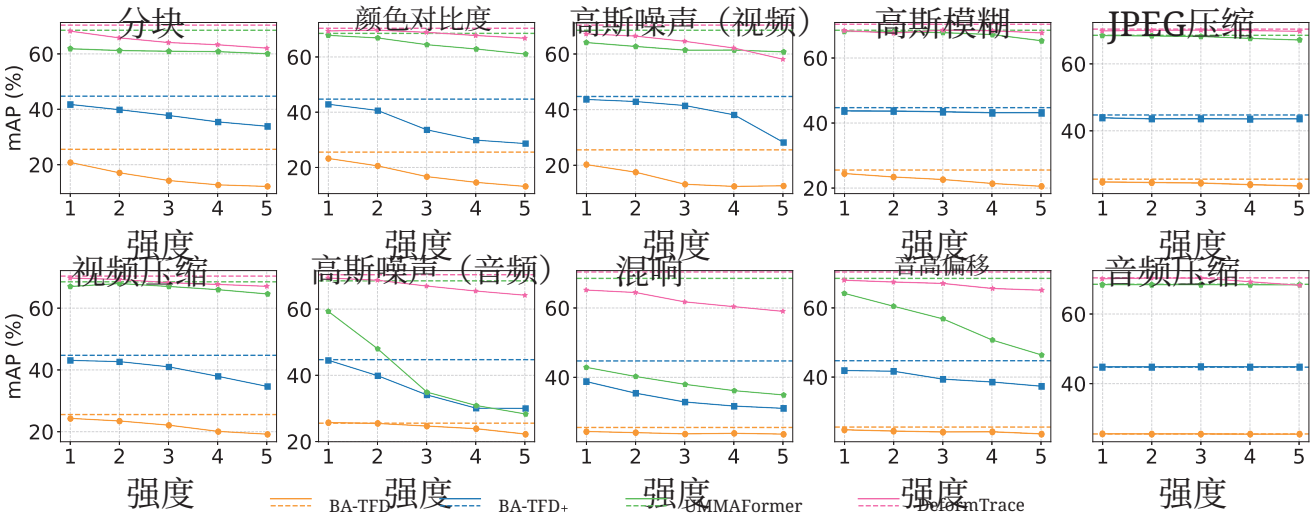


图3：在多种压缩与退化场景下的鲁棒性评估。实验包括6种视觉失真（分块、颜色对比度、高斯噪声（视频）、高斯模糊、JPEG压缩和视频压缩）以及4种音频失真（高斯噪声（音频）、混响、音高偏移和音频压缩）。图中颜色表示方法；实线表示不同强度下的平均精度均值，虚线表示干净视频上的平均精度均值。DeformTrace在干净视频上取得最高平均精度均值，并在各种失真场景中展现出较强的鲁棒性。

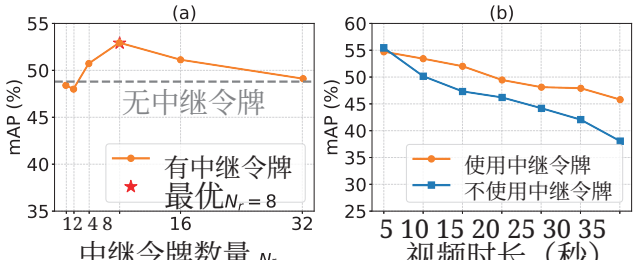


图4：（a）中继令牌数量的消融研究（平均视频长度：9秒）。（b）性能与视频持续时间的关系。

在没有中继令牌的情况下，因为每个子状态空间仍包含过多词元，因而无法缓解长程衰减。随着 N_r 增加，平均精度均值显著提升，并在 $N_r = 8$处达到峰值，此时每个子空间的令牌分布与解码器查询 N_q良好对齐。超过该点后，进一步增大 N_r 会因过度分割导致性能下降，从而阻碍全局信息流并提高复杂度。

性能与视频时长的关系。我们选取了 5 至 35 秒之间的七个特定时长，并从 AV-Deepfake1M 中抽取与每个目标时长最接近的 1,000 个视频，构建子数据集。通过这一设置，我们可以系统地考察模型性能如何随序列长度变化。如图 4(b) 所示，随着时长增加，采用中继令牌的模型优于其他模型，展现出更强的长程依赖捕捉能力。相比之下，不采用中继令牌的模型在短视频上的表现略好，但在较长序列上的性能会迅速下降。结果凸显了中继令牌对于维持随时间推移进行稳健时序推理的重要性。

鲁棒性研究

为评估模型的鲁棒性，我们采样了2,500个视频（2,000个来自LAV-DF，500个来自AV-Deepfake1M），并对每个视频施加10种常见传输扰动（6种视觉、4种音频），每种扰动设置5个强度级别，共得到125,000个降质视频。我们将模型与三种音视频TFL方法进行测试，并使用平均精度均值比较性能。如图3所示，DeformTrace在干净视频上取得最高平均精度均值，并对颜色对比度、高斯模糊、音高偏移以及三种压缩类型表现出较强鲁棒性。尽管在其他失真的高强度下性能有所下降，我们的方法仍持续优于基线方法，展现出更优的鲁棒性。

结论与讨论

我们提出DeformTrace，一种新颖的TFL框架，它融合状态空间模型（SSMs），以获得更优的性能-效率权衡。该框架包含具有动态感受野的可变形自状态空间模型（DS-SSM），用于精确边界定位；用于对稀疏伪造中的跨序列交互进行建模的可变形交叉状态空间模型（DC-SSM）；以及用于缓解长程信息衰减的中继令牌机制。通过将Transformer的全局建模与状态空间模型的效率相结合，DeformTrace以显著更快的推理实现精确定位，并在精确度与鲁棒性上均超越以往方法。

尽管在我们的工作中，DC-SSM仅用于建模伪造查询与特征之间的关系，但它也可以推广到任意两种独立序列，例如音频和视频，用于音视对应学习，从而拓展时序伪造定位之外的应用范围。

致谢

本研究部分受到国家自然科学基金委员会（62171326、62371350、62471343）、2025年新疆生产建设兵团重点科技研究项目（2025AB029）、广东OPPO移动通信有限公司和武汉大学超算中心的资助。

参考

Afouras, T.；Chung, J. S.；以及 Zisserman, A. 2018年。LRS3-TED：用于视觉语音识别的大规模数据集。arXiv:1809.00496。Ali, A.；Zimerman, I.；以及 Wolf, L. 2025年。Mamba模型的隐藏注意力。在计算语言学协会第*63*届年会论文集（*ACL*）中，1516–1534。Bai, Y.；Wang, Y.；Tong, Y.；Yang, Y.；Liu, Q.；以及 Liu, J. 2020年。用于时序动作提议生成的边界内容图神经网络。在欧洲计算机视觉会议论文集（*ECCV*）中，《计算机科学讲义》，121–137。Cham：Springer Nature Switzerland。Cai, Z.；Ghosh, S.；Adatia, A. P.；Hayat, M.；Dhall, A.；Gedeon, T.；以及 Stefanov, K. 2024年。AV-Deepfake1M：大规模大语言模型驱动的音视频深度伪造数据集。在第*32*届*ACM*国际多媒体会议论文集（*ACM MM*）中，7414–7423。Cai, Z.；Ghosh, S.；Dhall, A.；Gedeon, T.；Stefanov, K.；以及 Hayat, M. 2023年。Glitch in the Matrix：用于内容驱动的音视频伪造检测与定位的大规模基准。计算机视觉与图像理解，236：103818。Cai, Z.；Stefanov, K.；Dhall, A.；以及 Hayat, M. 2022年。你真的这样认为吗？内容驱动的音视频深度伪造数据集与用于时序伪造定位的多模态方法。在数字图像计算：技术与应用国际会议（*DICTA*）中，1–10。Carion, N.；Massa, F.；Synnaeve, G.；Usunier, N.；Kirillov, A.；以及 Zagoruyko, S. 2020年。使用Transformer的端到端目标检测。在欧洲计算机视觉会议论文集（*ECCV*）中，213–229。Springer。Chen, G.；Huang, Y.；Xu, J.；Pei, B.；Chen, Z.；Li, Z.；Wang, J.；Li, K.；Lu, T.；以及 Wang, L. 2024年。Video Mamba套件：作为视频理解通用替代方案的状态空间模型。arXiv:2403.09626。Cheng, F.；以及 Bertasius, G. 2022年。Tallformer：使用长记忆Transformer的时序动作定位。在欧洲计算机视觉会议论文集（*ECCV*）中，《计算机科学讲义》，503–521。Cham：Springer Nature Switzerland。Cheng, H.；Liu, H.；Cai, J.；以及 Ye, L. 2025年。CLFormer：用于时序伪造定位的跨语言Transformer框架。视觉智能，3(1)：13。Chung, J. S.；Nagrani, A.；以及 Zisserman, A. 2018年。VoxCeleb2：深度说话人识别。在*Interspeech 2018*论文集中，1086–1090。ISCA。

Dao, T.；以及 Gu, A. 2024年。Transformer即状态空间模型：通过结构化状态空间对偶实现的通用模型与高效算法。在第*41*届国际机器学习会议论文集（*ICML*）中。

Feng, C.; Chen, Z.; and Owens, A. 2023. 通过视听异常检测进行自监督视频取证。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，10491–10503。

Gu, A.; and Dao, T. 2024. Mamba：采用选择性状态空间的线性时间序列建模。在第*1*届国际语言建模会议（*ICLM*）论文集。Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.-T.; and Li, Y. 2025. MambaIRv2：注意力状态空间复原。在 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，28124–28133。Haliassos, A.; Ma, P.; Mira, R.; Petridis, S.; and Pantic, M. 2023. 从原始数据联合学习视觉与听觉语音表征。在 第*11*届国际学习表征会议（*ICLR*）论文集。Hatamizadeh, A.; and Kautz, J. 2025. MambaVision：混合式Mamba-Transformer视觉骨干网络。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，25261–25270。Hu, J.; Zheng, J.; Wei, J.; Zhang, J.; and Stiefelhagen, R. 2024. 面向大视场分割的可变形Mamba。arXiv:2411.16481。Huang, J.; Yuan, X.; Lam, C.-T.; Im, S.-K.; Lei, F.; and Bi, X. 2025. TransHFC：用于时序伪造定位的超图滤波卷积与Transformer框架。*IEEE*视频技术电路与系统汇刊。Kim, H.-J.; Hong, J.-H.; Kong, H.; and Lee, S.-W. 2024. TE-TAD：通过时间对齐坐标表达实现端到端时间动作检测。在 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，18837–18846。Koutlis, C.; and Papadopoulos, S. 2024. DiMoDif：用于视听深度伪造检测与定位的话语模态信息区分。arXiv:2411.10193。Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. 学习显著边界特征以实现无锚点时序动作定位。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，3320–3329。

Lin, T.; Liu, X.; Li, X.; Ding, E.; and Wen, S. 2019年。BMN：用于时序动作提议生成的边界匹配网络。载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，3889–3898。

Lin, T.; Zhao, X.; Su, H.; Wang, C.; and Yang, M. 2018年。BSN：用于时序动作提议生成的边界敏感网络。载于欧洲计算机视觉会议（*ECCV*）论文集，《计算机科学讲义》，3–21。Cham：Springer Nature Switzerland。

Liu, L.; Zhang, M.; Yin, J.; Liu, T.; Ji, W.; Piao, Y.; and Lu, H. 2025年。DefMamba：可变形视觉状态空间模型。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，8838–8847。

Liu, S.; Sui, L.; Zhang, C.-L.; Mu, F.; Zhao, C.; 和 Ghanem, B. 2024年a. 利用时间因果性实现先进的时间动作检测。arXiv:2407.17792。

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; 和 Bai, X. 2022. 基于Transformer的端到端时间动作检测。*IEEE*图像处理汇刊，31: 5427–5441。

Liu, Y.;Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; 和 Liu, Y. 2024年b. VMamba：视觉状态空间模型。神经信息处理系统进展，37: 103031–103063。

Long, F.; Yao, T.; Qiu, Z.; Tian, X.; Luo, J.; 和 Mei, T. 2019年. 用于动作定位的高斯时间感知网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），344–353。

Loshchilov, I.; and Hutter, F. 2019年。解耦权重衰减正则化。载于第七届国际学习表征会议（*ICLR*）论文集。

Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; and Bharaj, G. 2024年。AVFF：面向视频深度伪造检测的视听特征融合。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，27102–27112。

Patro,B. N.;andAgneeswaran, V.S. 2024年。SiMBA：面向视觉与多变量时间序列的简化版基于Mamba的架构。arXiv:2403.15360。

Pei, X.; Huang, T.; and Xu, C. 2025年。EfficientVMamba：面向轻量级视觉Mamba的空洞选择性扫描。载于*AAAI*人工智能会议（*AAAI*）论文集，第39卷，6443–6451。

´P´erez-Vieites, D.; Moreira-P´erez, J. J.; Arag´on-Kifute, A.; Rom´an-Sarmiento,R.; andCastro-Gonz´alez,R. 2024年。Vigo：视听伪造检测与片段定位。载于第*32*届*ACM*国际多媒体会议（*ACM MM*）论文集，11360–11364。

Smeu, S.;Boldisor, D.-A.; Oneata, D.;and Oneata, E. 2025年。利用无监督学习规避视听深度伪造检测数据集中的捷径。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，18815–18825。

Tan, J.; Tang, J.; Wang, L.; and Wu, G. 2021。用于直接动作提议生成的宽松型Transformer解码器。载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，13526–13535。

Wang, J.; Yan, J. N.; Gu, A.; and Rush, A. M. 2023. 无注意力预训练。载于计算语言学协会论文集：*EMNLP*，58–69。新加坡：计算语言学协会。

Wang, P.; Lu, S.; Dai, C.; Dai, S.; and Guo, B. 2025年。结合状态敏感型Mamba与质心序列增强的时序动作定位。*Neurocomputing*，620：129246。Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; and Ren, K. 2023。AVoiD-DF：用于检测深度伪造的视听联合学习。*IEEE Transactions on Information Forensics and Security*，18：2015–2029。Ye, Z.; Xia, K.; Fu, Y.; Dong, X.; Hong, J.; Yuan, X.; Diao, S.; Kautz, J.; Molchanov, P.; and Lin, Y. C. 2025年。Long-Mamba：通过免训练感受野扩展增强Mamba的长上下文能力。载于第*13*届国际学习表征会议（*ICLR*）论文集。

Yu, Y.; Liu, X.; Ni, R.; Yang, S.; Zhao, Y.; and Kot, A. C. 2023. PVASS-MDD：用于多模态深度伪造检测的预测式视听对齐自监督。*IEEE*视频技术电路与系统汇刊, 34: 6926–6936.

Zhang, C.; Wu, J.; and Li, Y. 2022. Actionformer：使用Transformer定位动作时刻。载于欧洲计算机视觉会议论文集（*ECCV*），《计算机科学讲义》，492–510。瑞士Cham：SpringerNature Switzerland.

Zhang, R.; Wang, H.; Du, M.; Liu, H.; Zhou, Y.; and Zeng, Q. 2023. UMMAFormer：用于时间伪造定位的通用多模态自适应Transformer框架。载于第*31*届*ACM*国际多媒体会议论文集（*ACM MM*），8749–8759.

Zhang, Y.; Miao, C.; Luo, M.; Li, J.; Deng, W.; Yao, W.; Li, Z.; Hu, B.; Feng, W.; Gong, T.; 等。2024年。MFMS：为深度伪造检测与定位任务学习模态融合特征和模态特定特征。载于第*32*届*ACM*国际多媒体会议论文集（*ACM MM*），11365–11369。

Zhou, Y.; 和 Lim, S.-N. 2021。联合音频-视觉深度伪造检测。载于*IEEE/CVF*国际计算机视觉会议论文集（*ICCV*），14780–14789。

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; 和 Wang, X. 2024年。视觉Mamba：利用双向状态空间模型进行高效视觉表征学习。载于第*41*届国际机器学习会议（*ICML*）论文集。

Zhu, X.； Wang, S.； Yang, J.； Yang, Y.； Tu, W.； 和 Wang, Z. 2025年。采用寄存器增强表征学习的基于查询的音频-视觉时序伪造定位。发表于第*33*届*ACM*国际多媒体会议（*ACM MM*）论文集，8547–8556。