arXiv:2601.01784v1 [cs.CV] 5 Jan 2026

DDNet：用于时序伪造定位的双流图学习与解耦框架

Boyang Zhao[1], Xin Liao[1*], Jiaxin Chen[2],Xiaoshuai Wu[1],和 Yufeng Wu[1]

[1]湖南大学网络空间安全与技术学院，中国长沙[2]长沙理工大学计算机科学与技术学院，中国长沙

摘要—人工智能生成内容技术的迅速发展，使得只需篡改视频中的少量片段便能误导观众，导致视频级检测不够准确且缺乏说服力。因此，旨在精确定位篡改片段的时序伪造定位（TFL）变得至关重要。然而，现有方法往往受限于局部视图，无法捕捉全局异常。为此，我们提出了一种用于时序伪造定位的双流图学习与解耦框架（DDNet）。DDNet通过协同运用针对局部伪影的时间距离流 和用于建模长程连接的语义内容流，避免全局线索被局部平滑性淹没。此外，我们提出痕迹解耦与适应（TDA），以分离通用伪造指纹；同时引入跨层级特征嵌入（CLFE），通过深度融合层次特征构建稳健的特征基础。在ForgeryNet和TVIL基准测试上的实验表明，我们的方法在AP@0.95上比最先进的方法高出约9%，跨域鲁棒性也显著提升。

索引词—**时序伪造定位、双流图学习、局部-全局推理**

I. 引言

AIGC 技术的快速发展使视频伪造日益复杂且难以察觉 [1], [2],, 攻击者可通过篡改微小片段来造成恶意语义扭曲或误导性效果[3]。在伪造痕迹极其细微的情况下，难以获得准确的二元分类结果；此外，由于缺乏可解释性，这些结果往往无法提供令人信服的证据 [4], [5]。因此，时序伪造定位（TFL）已成为视频取证中的一项关键任务 [5], [6], ，因为它能够提供篡改片段的精确时间戳，从而支持可解释的证据分析，并显著减少人工检查

工作（如图1(a)所示）。

然而，现有的 TFL 范式主要依赖标准卷积神经网络或 Transformer，但在伪造检测场景中，两者都存在局部视图的局限（如图 1(b) 所示）。卷积神经网络受限于固定的感受野 [3],[6]–[8], ，难以参考相距较远的真实帧进行验证。相反，标准 Transformer [5],[9]–[11] 往往会使注意力权重偏向高度相似的

通讯作者：xinliao@hnu.edu.cn

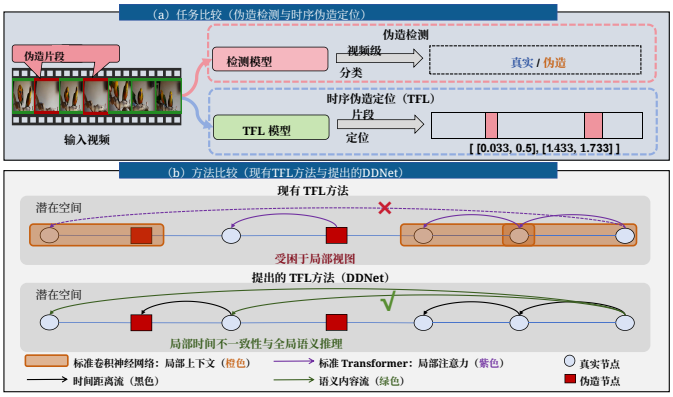


图1. 研究背景与动机示意图。(a) 伪造检测与时序伪造定位（TFL）的比较。伪造检测仅输出单个视频级标签（真实/伪造），往往会忽略短暂的伪造片段。相比之下，TFL旨在精确确定伪造片段的位置。(b) 现有TFL方法与所提出的DDNet的比较。现有TFL方法往往受局部视图限制，无法捕获相距较远节点之间的长程依赖。我们提出的DDNet通过引入双流图学习（DSGL）克服了这一问题。通过协同时序距离流和语义内容流，我们将局部不一致性与全局线索相结合，从而实现更精准的定位。

局部邻域，从而难以捕捉关键的全局不一致性。因此，细微的伪造痕迹常被占主导地位的真实背景所掩盖，导致定位不准确。

为利用全局上下文定位伪造片段，我们提出了 DDNet，一种双流图学习与解耦框架。如图2所示，我们的框架由三个主要组件构成。首先，鉴于伪造痕迹具有隐蔽性，我们引入跨层特征嵌入（CLFE），以提供稳健的特征基础。通过融合 CLIP [12] 中的高层语义与 ResNet [13], 中的低层纹理，该模块能够捕获检测细微篡改所必需的多粒度线索。在此基础上，为解决局部视图的核心局限，我们设计了双流图学习（DSGL）模块。该组件整合了时间距离流以捕获局部不一致性，并通过语义内容流建立远距离帧之间的全局语义连接。通过显式连接观测结果与远距离的真实锚点，DSGL 缓解了局部注意力机制固有的过平滑现象，确保细微的伪造痕迹不会被其邻近信息抑制。

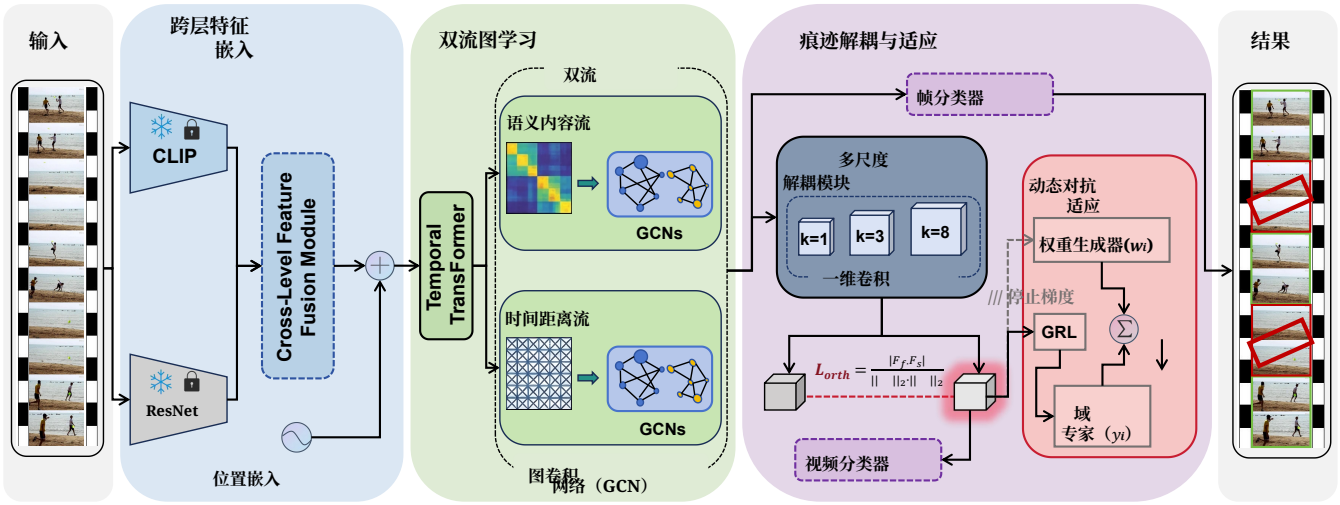


图2. 所提出的 DDNet 框架概览。该框架由三个主要组件构成：(a) **跨层级特征嵌入（CLFE）：** 该模块使用冻结的 CLIP 和 ResNet 编码器提取视觉特征，随后通过双向交叉注意力机制，将语义先验与纹理细节融合，以弥合特征差距。(b) **双流图学习（DSGL）：** 该核心模块旨在克服局部视图的局限。它整合了时间距离流以捕捉局部时间不一致性，并整合语义内容流，以对彼此分离的篡改片段进行全局语义推理。(c) **痕迹解耦与适应（TDA）：** 该辅助训练模块施加对抗约束和正交约束，以从领域特定噪声中分离通用伪造指纹。最后，将学习到的表征投影后生成帧级篡改概率。

点，DSGL 缓解了局部注意力机制固有的过平滑现象，确保细微的伪造痕迹不会被其邻近信息抑制。最后，为确保推理聚焦于内在伪造痕迹，我们提出了痕迹解耦与自适应（TDA）。作为一种隐式约束，该辅助模块促使模型丢弃无关噪声，聚焦于通用伪造指纹，从而增强其在跨域场景中的鲁棒性。

总而言之，我们的主要贡献如下：

- 我们提出了 DDNet，这是一种采用双流图学习（DSGL）模块的框架。通过明确协同建模局部时间不一致性与全局语义推理，该框架有效突破了局部视图的局限，并显著提高了定位精度。
- 我们提出了跨层级特征嵌入（CLFE）模块和辅助性的痕迹解耦与适应（TDA）模块。通过捕获跨层级互补特征，并施加隐式约束以关注通用伪造指纹，这些组件协同提升了模型的鲁棒性。
- 在 ForgeryNet 和 TVIL 基准测试上进行的大量实验表明，DDNet 达到了最先进水平，性能优于现有方法，AP@0.95 提升约 9%，同时跨域鲁棒性也显著增强。

II. 相关工作

A. 视频伪造检测

早期的视频伪造检测研究主要将问题表述为二元分类任务。最初的方法专注于识别单帧中的图像级伪影；后续方法则利用视频片段中的时间不一致性。尽管这些方法在 FaceForensics 和 DFDC 等基准测试上取得了令人瞩目的准确率，但它们只能提供粗粒度标签（真实与伪造）。正如近期研究所指出的，二元分类缺乏视频取证所需的可解释性，因为它无法指出具体的篡改片段，也无法提供令人信服的证据。这一局限促使研究转向更细粒度的任务：时序伪造定位（TFL）。

B. 时序伪造定位

TFL 旨在精确定位视频中的篡改片段。这一目标与时间动作定位（TAL）十分相似，后者侧重于在未剪辑视频中定位动作实例。受这种相似性的推动，TFL 方法主要通过改造 TAL 架构发展而来，主要分为两种范式：边界匹配 [3], [6] 和单阶段回归 [5], [9], [10]。

早期基于CNN的改进方法 [3], [6]受限于固定的感受野，难以捕捉长程依赖。尽管近期的基于Transformer的方法 [5],[9], [10] 有所改进，但仍受局部视图限制。其注意力机制往往过度强调相邻节点，导致关键的全局不一致性线索被局部上下文淹没。

此外，图卷积网络 [18]和特征解耦等先进技术在 TFL 中仍未得到充分探索。现有方法大多依赖纠缠特征，导致语义过拟合。为弥补这一不足，DDNet 集成了双流图学习（DSGL）以进行全局推理，并采用轨迹解耦与适配（TDA）来分离通用伪造

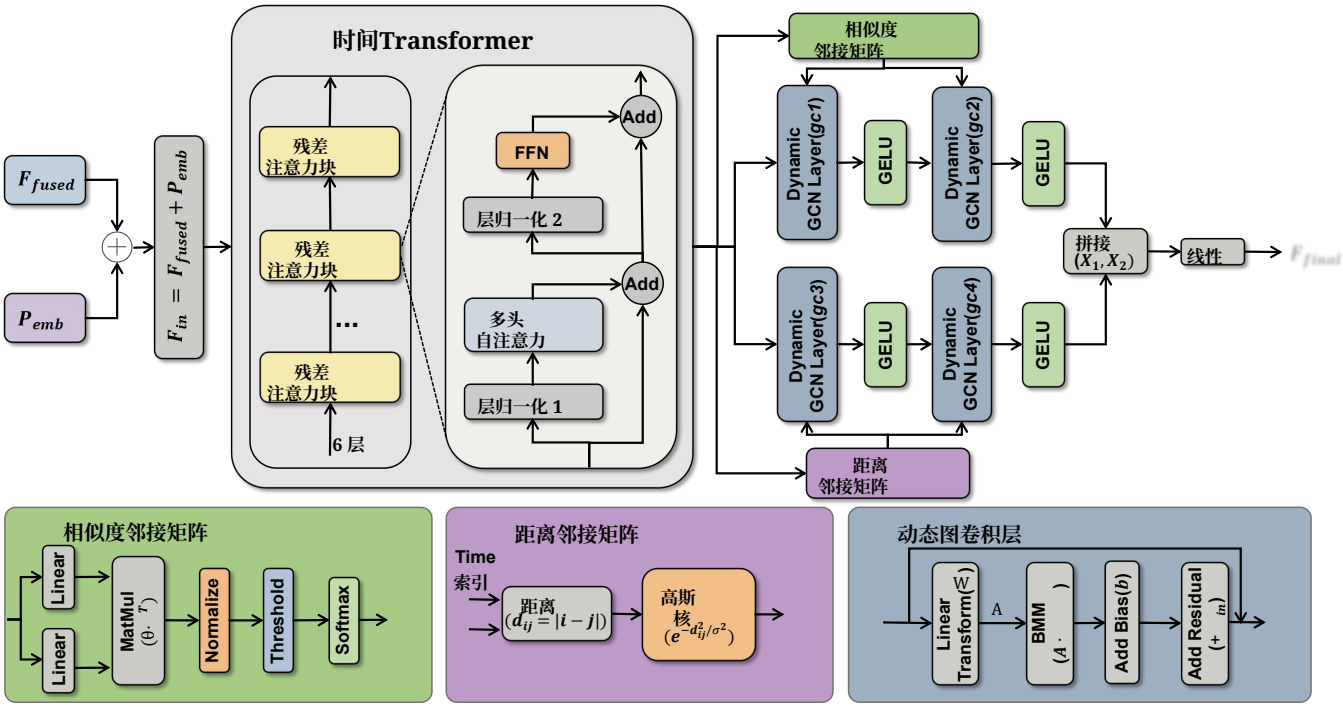


图3. 所提出的双流图学习（DSGL）模块示意图。

指纹，有效克服了局部视图和泛化能力较弱的局限。

III. 方法论

A. 概述

如图2所示，面向时间伪造定位的双流图学习与解耦框架（DDNet）建立了一个由三个主要组件构成的统一端到端框架：

1) **跨层级特征嵌入（CLFE）**：通过利用视觉编码器进行层次化提取，并借助堆叠式双向交叉注意力机制深度融合特征，从而构建稳健的特征基础。

2) **双流图学习（DSGL）**：利用双流图架构的核心推理模块。该模块协调时间距离流以强化局部不一致性，并协调语义内容流以捕获全局依赖关系，从而细化帧级表示。

3) **痕迹解耦与适应（TDA）**：采用动态混合专家（MoE）对抗机制的辅助正则化模块。该模块通过解耦与任务无关的噪声，促使骨干网络学习域不变伪造指纹。

至关重要的是，该框架采用联合优化，并直接根据DSGL 输出预测帧级篡改概率 $\mathbf{P} = \{p_t\}_{t=1}^{T}$ 。

B. 跨层级特征嵌入（*CLFE*）

为捕捉细微的纹理伪影和高层语义异常，我们采用冻结的 CLIP [12] 编码器和 ResNet [13]编码器，分别提取语义特征 $\mathbf{F}_{CLIP}$和纹理特征 $\mathbf{F}_{Res}$。这些异构特征被映射到统一维度 D，并通过堆叠式双向交叉注意力模块进行融合。

在每个模块中，令$\mathbf{H}_{CLIP}$ 和$\mathbf{H}_{Res}$ 表示两个中间特征表示。这些表示通过交换信息进行交互，其中一个作为查询，从另一个表示中检索上下文（后者作为键和值）。形式化地，跨模态交互定义如下：

$$\begin{aligned}\mathbf{H}'_{CLIP} &= \mathrm{MHA}(\mathrm{LN}(\mathbf{H}_{CLIP}), \mathrm{LN}(\mathbf{H}_{Res}), \mathrm{LN}(\mathbf{H}_{Res})),\\ \mathbf{H}'_{Res} &= \mathrm{MHA}(\mathrm{LN}(\mathbf{H}_{Res}), \mathrm{LN}(\mathbf{H}_{CLIP}), \mathrm{LN}(\mathbf{H}_{CLIP}))\end{aligned} \tag{1}$$

其中，MHA(·) 表示多头注意力，LN(·)表示层归一化，而 $\mathbf{H}'$ 表示更新后的特征。这种双向机制确保了语义与纹理的深度融合。精炼后的特征经过拼接和投影，生成融合表示$\mathbf{F}_{fused} \in \mathbb{R}^{T\times D}$。

最后，为了保留时间顺序，我们引入可学习的位置嵌入$\mathbf{P}_{emb} \in \mathbb{R}^{T\times D}$，其参数由帧索引 $t \in \{0, \ldots, T-1\}$确定。在序列输入时间Transformer之前，这些嵌入会逐元素加到$\mathbf{F}_{fused}$ 上。

C. 双流图学习（*DSGL*）

为了解决局部视图的局限性，我们引入了 DSGL 模块（图 3）。首先，将融合特征组合

表I 在 ForgeryNet 和 TVIL 基准测试上的性能比较。**最佳** 表示最佳性能。

方法	ForgeryNet（标准）			ForgeryNet（均衡）			TVIL（图像修复）		
	@0.5	@0.75	@0.95	@0.5	@0.75	@0.95	@0.5	@0.75	@0.95
Tridet [19]	-	-	-	21.01	2.63	0.03	35.58	6.24	0.43
DCAN [20]	71.59	62.02	32.08	62.80	50.67	22.91	79.64	51.59	0.25
ActionFormer [11]	84.49	78.74	66.10	72.30	64.90	54.00	86.30	83.00	28.20
UMMAFormer [5]	82.30	78.90	67.70	74.71	67.43	54.32	**88.70**	**84.70**	62.40
Ours	**87.59**	**83.46**	**76.70**	**74.89**	**68.81**	**57.42**	81.64	78.25	**70.81**

位置嵌入$\mathbf{P}_{emb}$后，输入**时间Transformer**（6 层）进行处理，以编码序列依赖关系。输出特征作为两个并行图流的共享输入，用于捕获互补线索。

1）时间距离流。为减轻预测闪烁，该流利用静态图 A_{dist} 来施加局部不一致性（层 gc3、gc4）。如图 3 所示，距离邻接矩阵 $\mathbf{A}_{dist} \in \mathbb{R}^{T\times T}$ 通过高斯核，根据帧 i 和帧 j之间的时间间隔 $|i-j|$计算得出：

$$\mathbf{A}_{dist}(i,j) = \exp\left(-\frac{|i-j|^2}{2\sigma^2}\right), \tag{2}$$

其中，σ 控制局部感受野的带宽，起到平滑滤波器的作用。

2）语义内容流。为捕捉长程语义依赖（层 gc1、gc2），我们基于特征相似度构建动态图。如图 3 所示，输入特征经线性层投影后计算成对相似度，随后进行归一化和硬阈值化：

$$\mathbf{A}_{sim} = \text{Softmax}\left(\text{Threshold}_{\tau}\left(\text{Norm}\left(\mathbf{X}\mathbf{W}_{\theta}(\mathbf{X}\mathbf{W}_{\phi})^{T}\right)\right)\right), \tag{3}$$

其中，$\mathbf{W}_{\theta}$，$\mathbf{W}_{\phi}$ 为投影矩阵，τ 是用于消除噪声的阈值。根据补充材料中的敏感性分析，为在大规模基准测试中获得稳健的性能，我们通过经验将 $\tau = 0.7$ 设为该值。

3）动态 GCN 与融合（图3）。两个流均采用图3所示的动态图卷积层。对于输入$\mathbf{H}_{in}$ 和邻接矩阵$\mathbf{A}$，输出$\mathbf{H}_{out}$定义为：

$$\mathbf{H}_{out} = \mathbf{A}(\mathbf{H}_{in}\mathbf{W}) + \mathbf{b} + \mathbf{H}_{in}, \tag{4}$$

其中，$\mathbf{W}$ 和$\mathbf{b}$ 分别表示权重和偏置。最后，将两个流的特征拼接起来，并通过线性投影进行融合，得到$\mathbf{F}_{final}$。

D. 痕迹解耦与适应（*TDA*）

为从领域特定噪声中分离通用伪造指纹，我们引入了 TDA 模块（图 2），该模块包含两个关键组件。

1）多尺度解耦。我们首先通过平均池化将帧特征聚合为视频级原型。随后，利用多尺度一维卷积处理该原型（$k \in \{1,3,8\}$），以捕获多尺度伪造模式。接着，将这些特征投影为两个正交向量：通用伪造特征$\mathbf{F}_f$ 和无关特征$\mathbf{F}_s$。为确保二者分离，我们最小化它们的余弦相似度：

$$\mathcal{L}_{orth} = \frac{1}{B}\sum_{i=1}^{B}\left|\frac{\mathbf{F}_f^{(i)}\cdot\mathbf{F}_s^{(i)}}{\|\mathbf{F}_f^{(i)}\|\|\mathbf{F}_s^{(i)}\|}\right|, \tag{5}$$

其中，B为批大小，$\|\cdot\|$ 表示 L_2 范数。

2）动态对抗适应。为实现鲁棒泛化，将$\mathbf{F}_f$ 输入用于领域适应。我们不采用单个判别器，而是使用动态集成。首先，**权重生成器**（见图2）预测领域相关性分数 $\boldsymbol{\omega} \in \mathbb{R}^K$（其中 K 为领域数量）。关键在于，应用停止梯度操作 sg(·)，以稳定生成器：

$$\boldsymbol{\omega} = \text{Softmax}(\text{Linear}(\text{sg}(\mathbf{F}_f))). \tag{6}$$

同时，$\mathbf{F}_f$ 经过梯度反转层（GRL）[21] 后，输入由**领域专家** $\{\mathcal{D}_k\}_{k=1}^{K}$ 组成的专家组。最终的对抗决策是各专家输出的加权求和：

$$O_{adv} = \sum_{k=1}^{K}\omega_k \cdot \mathcal{D}_k(\text{GRL}(\mathbf{F}_f)). \tag{7}$$

E. 优化目标

该框架以端到端方式进行优化。TDA模块作为辅助正则化组件，引导骨干网络学习鲁棒表征。总目标函数定义如下：

$$\mathcal{L}_{total} = \mathcal{L}_{frame} + \lambda_{vid}\mathcal{L}_{video} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{orth}\mathcal{L}_{orth}, \tag{8}$$

其中，$\mathcal{L}_{frame}$ 和 $\mathcal{L}_{video}$ 分别表示用于帧级定位和视频级检测的二元交叉熵损失。具体而言，$\mathcal{L}_{adv}$ 是聚合后的对抗输出 O_{adv} 与真实伪造领域标签之间计算得到的交叉熵损失。

IV. 实验

A. 实验设置

数据集与指标。 我们在两个基准测试上评估 DDNet：（1）**ForgeryNet** [22],一个包含多种操纵技术的大规模 TFL 数据集。我们构建了标准 和

表II 在 ForgeryNet 和 TVIL 基准测试上的性能比较。**BOLD** 表示最佳性能。

方法	ForgeryNet（标准） → TVIL			ForgeryNet（均衡） → TVIL			TVIL → ForgeryNet（均衡）		
	@0.5	@0.75	@0.95	@0.5	@0.75	@0.95	@0.5	@0.75	@0.95
Tridet [19]	-	-	-	1.05	0.03	0	0.56	0	0
DCAN [20]	1.49	0.43	0	0.83	0.21	0	1.12	0.27	0
ActionFormer [11]	2.76	1.40	0.02	1.12	0.33	0	0.82	0.10	0
UMMAFormer [5]	3.02	1.80	0	2.15	0.98	0	1.26	0.18	0
我们的方法（无_TDA）	7.25	5.27	0.25	4.25	3.14	0.02	**1.51**	**0.56**	0
Ours	**11.72**	**8.25**	**0.62**	**7.64**	**6.43**	**0.09**	1.33	0.52	**0**

表III 核心组件消融研究。

输入		组件			指标		
CLIP	ResNet	CLFE	GCN	TDA	@0.5	@0.75	@0.95
✓		-	✓	✓	47.00	35.04	22.84
	✓	-	✓	✓	58.52	48.90	26.00
✓	✓		✓	✓	61.85	49.64	26.75
✓	✓	✓		✓	65.39	59.20	51.10
✓	✓	✓	✓		71.57	63.82	53.18
✓	✓	✓	✓	✓	**74.89**	**68.81**	**57.42**

均衡 子集，分别用于评估总体性能和类别均衡性能（详情请参见**补充材料**）。（2）**TVIL** [5], 一个专用于视频修复场景的基准测试，通过使用先进的修复算法在随机时间区间移除物体构建而成。

评估指标。我们报告 tIoU阈值 $\{0.5, 0.75, 0.95\}$ 下的平均精度（AP）。我们特别关注 AP@0.95，因为它反映了模型执行高精度定位的能力。

基线方法。我们将模型与通用 TAL 模型（Ac- tionFormer [11], DCAN [20]）以及专用 TFL 框架（UMMAFormer [5]）进行比较。值得注意的是，VIGO [9] 与 UMMAFormer 共用视觉架构；因此，与 UMMAFormer 的比较反映了当前基于视觉的方法所达到的最先进水平。

实现细节。我们的模型使用 PyTorch 实现，并采用 AdamW优化器进行训练，初始学习率为 2×10^{-5}。批大小设为 64，时间维度固定为 $T = 512$。我们使用冻结的 CLIP 和 ResNet-50 作为特征提取器。对于损失函数，平衡系数设为 $\lambda_{vid} = 0.3$、$\lambda_{adv} = 0.005$ 和 $\lambda_{orth} = 1.0$。

B. 域内性能分析

表I总结了定量结果。在**ForgeryNet** 标准子集上，DDNet的AP@0.95达到76.70%，比最先进水平的UMMAFormer高出**9.0%**。同样，在包含具有挑战性的图像修复攻击的**TVIL**上，DDNet仍保持更高的精确率（70.81%，而UMMAFormer为62.40%）。在严格的时间交并比阈值下取得如此显著的优势，验证了我们的DSGL模块在结构上的优势。不同于受局部邻居限制的标准Transformer，DSGL将时间距离不一致性与语义内容推理相结合，有效提升了定位精度。

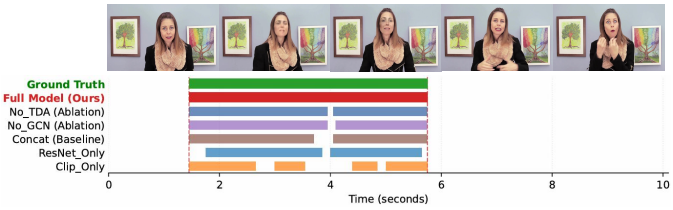


图4. **定性可视化**。真实标注（绿色）显示一段连续的伪造内容。消融变体存在严重的预测碎片化问题。我们的完整模型（红色）能够保持一致性，避免消融变体中出现的间断。

C. 跨域泛化

表II揭示了一种不对称的泛化模式。

从ForgeryNet迁移到TVIL。尽管存在语义差异，DDNet的AP@0.5仍达到11.72%，是UMMAFormer的**4倍**。这一成果表明，ForgeryNet中多样的篡改痕迹有效涵盖了TVIL中较简单的图像修复痕迹。得益于TDA模块，我们的编码器能够将内在的伪造指纹与无关噪声解耦，从而轻松泛化到TVIL。

从 TVIL 迁移到 ForgeryNet。相反，反向迁移则颇具挑战。我们将此归因于领域过拟合：在 TVIL 上训练的模型只关注特定的图像修复伪影。由此产生的表征对于人脸取证并不完整，因为它们缺乏复杂深度伪造检测所需的高级语义和生物学线索（例如面部不一致性）。

D. 消融研究

我们在**ForgeryNet**均衡子集上分析各组件的贡献，见表III。

1) 图推理（DSGL）的影响。为单独评估图卷积网络的贡献，我们移除并行的图卷积网络流（第4行），同时保留时间Transformer。请注意，移除整个 DSGL 将导致时间建模能力丧失。结果显示，平均精度@0.95 约下降6%，这证实了仅依赖时间注意力不足以实现精确定位。图卷积网络流对于显式强化局部不一致性和全局线索至关重要。

2）TDA 的有效性。去除辅助 TDA 模块（第 5 行）会导致域内性能略有下降，同时跨域泛化能力也有所减弱

（见表 II）。这验证了 TDA 能防止模型对特定视频内容过拟合，促使其学习能够更好地泛化到新域的通用伪造痕迹。

3）CLFE 的必要性。第 1–3 行表明，无论是单模态骨干网络（缺乏互补线索），还是朴素拼接（受特征异质性限制），都无法取得最优结果。相比之下，引入 CLFE 带来了显著的性能提升。这证明，我们的双向交叉注意力机制对于弥合语义先验与纹理伪影之间的差距不可或缺。

E. 定性分析

图4展示了一个连续伪造样本的检测结果。

变体中的碎片化。如图4所示，真实标注（绿色）表示一个连续的篡改片段。然而，消融变体存在严重的预测碎片化问题。具体而言，移除图推理（无_图卷积网络，紫色）或仅依赖单一模态（*Clip_*仅，橙色）都会产生不连续的片段。这种无法保持时间一致性的缺陷凸显了完整模型的必要性。

DDNet 的精确率。相比之下，我们的完整模型（红色）生成了高度连贯的候选片段，与真实标注精确对齐。DDNet 有效弥合了不完整变体中观察到的过渡间隙，消除了碎片化，并确保了更高的定位精度。

V. 结论

本文提出了 DDNet，显著提升了 TFL 中的定位精度和泛化能力。通过将局部不一致性与全局推理相结合，我们的 DSGL 模块有效克服了局部视图的局限性。结合 TDA 和 CLFE，该框架能够学习稳健的篡改指纹，并在不同领域间有效迁移。在 ForgeryNet 和 TVIL 上进行的大量实验表明，该方法性能达到最先进水平，尤其是在高精确率场景下。未来的工作将探索多模态集成和大型视觉模型（LVM），以弥合尚存的跨领域语义差距。

参考文献

[1] Andreas Rossler、Davide Cozzolino、Luisa Verdoliva、Christian Riess、Justus Thies 和 Matthias Nießner，“Faceforensics++：学习检测经过篡改的人脸图像”，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第1–11页。[2] Zhiyuan Yan、Taiping Yao、Shen Chen、Yandan Zhao、Xinghe Fu、Junwei Zhu、Donghao Luo、Chengjie Wang、Shouhong Ding、Yunsheng Wu 等，“Df40：迈向新一代深度伪造检测”，载于神经信息处理系统进展，第37卷，29434，2024年29387页–。[3] Zhixi Cai、Shreya Ghosh、Abhinav Dhall、Tom Gedeon、Kalin Stefanov 和 Munawar Hayat，“矩阵中的故障：用于内容驱动音视频伪造检测与定位的大规模基准测试”，计算机视觉与图像理解，第236卷，第103818页，2023年。[4] Changtao Miao、Yi Zhang、Weize Gao、Man Luo、Weiwei Feng、Zhiya Tan、Jianshu Li、Ajian Liu、Yunfeng Diao、Qi Chu 等，“Ddl：面向真实场景中可解释深度伪造检测与定位的数据集”，*arXiv*预印本 *arXiv:2506.23292*，2025年。

[5] Rui Zhang、Hongxia Wang、Mingshan Du、Hanqing Liu、Yang Zhou 和 Qiang Zeng，《Ummaformer：用于时序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，2023年，第8749–8759页。[6] Zhixi Cai、Kalin Stefanov、Abhinav Dhall 和 Munawar Hayat，《你真的这么想吗？由内容驱动的视听深度伪造数据集及用于时序伪造定位的多模态方法》，载于*2022*年国际数字图像计算：技术与应用会议（*DICTA*）。IEEE，2022年，第1–10页。[7] Tianwei Lin、Xiao Liu、Xin Li、Errui Ding 和 Shilei Wen，《BMN：用于时间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论文集，2019年，第3889–3898页。[8] Haisheng Su、Weihao Gan、Wei Wu、Yu Qiao 和 Junjie Yan，《BSN++：用于时间动作提议生成的尺度均衡关系建模互补边界回归器》，载于*AAAI*人工智能会议论文集，2021年，第35卷，第2602–2610页。[9] Diego Pérez-Vieites、Juan José Moreira-Pérez、Ángel Aragón-Kifute、Raquel Román-Sarmiento 和 Rubén Castro-González，《VIGO：视听伪造检测与片段定位》，载于第*32*届*ACM*国际多媒体会议论文集，2024年，第11360–11364页。[10] Haonan Cheng、Hanyue Liu、Juanjuan Cai 和 Long Ye，《CLFormer：用于时序伪造定位的跨语言Transformer框架》，*Visual Intelligence*，第3卷，第1期，第13页，2025年。[11] Chen-Lin Zhang、Jianxin Wu 和 Yin Li，《ActionFormer：使用Transformer定位动作时刻》，载于欧洲计算机视觉会议。Springer，2022年，第492–510页。[12] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等，《通过自然语言监督学习可迁移视觉模型》，载于国际机器学习会议。PMLR，2021年，第8748–8763页。[13] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，《用于图像识别的深度残差学习》，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第770–778页。[14] Lingzhi Li、Jianmin Bao、Ting Zhang、Hao Yang、Dong Chen、Fang Wen 和 Baining Guo，《用于更通用人脸伪造检测的Face X-ray》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2020年，第5001–5010页。[15] Zhihao Gu、Yang Chen、Taiping Yao、Shouhong Ding、Jilin Li 和 Lizhuang Ma，《深入局部：用于深度伪造视频检测的动态不一致性学习》，载于*AAAI*人工智能会议论文集，2022年，第36卷，第744–752页。[16] Yunzhuo Chen、Naveed Akhtar、Nur Al Hasan Haldar 和 Ajmal Mian，《结合时空一致性与注意力的深度伪造检测》，载于*2022*年国际数字图像计算：技术与应用会议（*DICTA*）。IEEE，2022年，第1–8页。[17] Brian Dolhansky、Joanna Bitton、Ben Pflaum、Jikuo Lu、Russ Howes、Menglin Wang 和 Cristian Canton Ferrer，《深度伪造检测挑战赛（DFDC）数据集》，*arXiv*预印本 *arXiv:2006.07397*，2020年。[18] Sijie Yan、Yuanjun Xiong 和 Dahua Lin，《用于基于骨架的动作识别的时空图卷积网络》，载于*AAAI*人工智能会议论文集，2018年，第32卷。[19] Dingfeng Shi、Yujie Zhong、Qiong Cao、Lin Ma、Jia Li 和 Dacheng Tao，《Tridet：采用相对边界建模的时间动作检测》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2023年，第18857–18866页。[20] Guo Chen、Yin-Dong Zheng、Limin Wang 和 Tong Lu，《DCAN：通过双重上下文聚合改进时间动作检测》，载于*AAAI*人工智能会议论文集，2022年，第36卷，第248–257页。[21] Yaroslav Ganin 和 Victor Lempitsky，《通过反向传播进行无监督领域适应》，载于国际机器学习会议。PMLR，2015年，第1180–1189页。[22] Yinan He、Bei Gan、Siyu Chen、Yichun Zhou、Guojun Yin、Luchuan Song、Lu Sheng、Jing Shao 和 Ziwei Liu，《ForgeryNet：用于全面伪造分析的多用途基准测试》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第4360–4369页。

DDNet：用于时序伪造定位的双流图学习与解耦框架（补充材料）

Boyang Zhao[1], Xin Liao[1*]、Jiaxin Chen[2], Xiaoshuai Wu[1],Yufeng Wu[1],和匿名ICME投稿[2]

[1]湖南大学网络空间安全与技术学院，中国长沙[2]长沙理工大学计算机科学与技术学院，中国长沙

在这份补充材料中，我们提供了因篇幅限制而未在正文中呈现的更多实验细节与分析。具体而言，我们首先详细介绍 ForgeryNet 数据集中所用标准和均衡子集的具体构成与分布（第 VI 节）。此外，我们还对双流图学习（DSGL）模块中的超参数 τ 进行了全面的敏感性分析，以验证我们的经验性参数选择（第 VII 节）。

VI. 数据集详情

在实验中，我们使用两个不同的评估子集评估 ForgeryNet [22] 数据集上的性能：标准子集和均衡子集。这些评估方案旨在考察模型在不同数据分布下的性能。

ForgeryNet [22] 是一个综合性数据集，具有丰富多样的操纵技术和庞大的规模。为了全面评估我们的方法，我们基于该数据集构建了两个不同的子集：

- **标准子集（自然分布）：** 首先，我们从原始数据集中随机抽样，构建了一个大规模子集。该子集包含 **59,994** 个训练视频、**14,019** 个测试视频和**14,020** 个验证视频。如图 5 所示，训练集中各类篡改方法的分布高度倾斜（呈长尾分布）。**具体而言，方法编号 0 表示真实视频，其余编号对应由不同操纵技术生成的视频**。这反映了现实场景中常见的自然不均衡现象。

- **平衡子集（均匀分布）：** 考虑到实验的计算效率，以及标准子集中长尾分布可能引入的偏置，我们进一步构建了一个平衡子集。该子集包含 **5,600** 个训练视频、**700**个测试视频和 **700** 个验证视频。与标准子集相比，我们严格确保**每种伪造方法在所有划分中均由相同数量的视频表示**。这种设置有助于公平评估模型在不同伪造技术上的泛化能力，而不会受到多数类的主导。

多种伪造技术，而不受多数类主导。

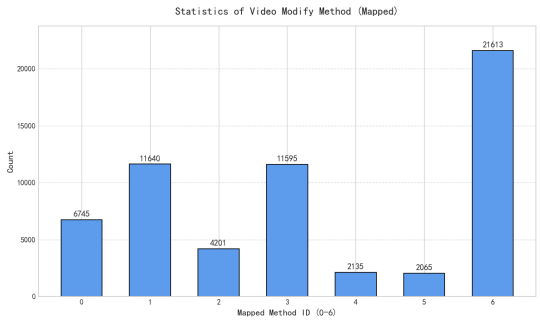


图 5. **标准子集**的**训练集**中各类篡改方法的分布。分布高度不均衡，其中方法 6 占据了数据的很大比例，而其他方法则代表不足。

VII. 阈值敏感性分析 τ

A. 超参数定义与评估指标

在我们提出的 DSGL模块的语义内容流 中，我们构建了一个动态图，以捕获长程语义依赖关系。超参数 τ 作为施加于归一化特征相似度矩阵的硬阈值。

- **含义 τ：** 该参数明确控制语义图的稀疏性。较低的 τ 允许保留更多边，这可能引入无关或嘈杂的连接。相反，较高的 τ 会实施更严格的筛选，仅保留最可靠的语义相关性，但可能会切断有用的长程依赖关系。

- **评估指标（平均精度均值）：** 为全面评估定位精度，我们报告平均精度均值（mAP）。具体而言，该指标是三个不同交并比（IoU）阈值下 AP 分数的平均值：

$$\mathrm{mAP} = \frac{\mathrm{AP@0.5} + \mathrm{AP@0.75} + \mathrm{AP@0.95}}{3} \tag{9}$$

通讯作者：xinliao@hnu.edu.cn

B. 结果与讨论

为了确定 τ 的最佳值，我们通过在 $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ 范围内改变 τ，评估模型在 *ForgeryNet* 均衡子集上的性能。定量结果如图 6 所示。

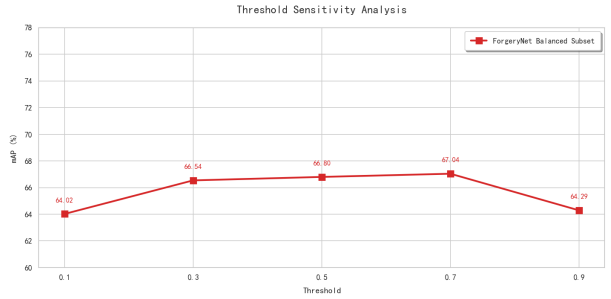


图 6. τ的敏感性分析。曲线展示了噪声抑制与语义保留之间的权衡。平均精度均值按 AP@0.5、AP@0.75 和 AP@0.95 的平均值计算。

权衡分析： 随着 τ 从 0.1 增加到 0.7，性能持续提升，并在**67.04%**时达到峰值。这一趋势表明，初始特征相似度矩阵中包含大量噪声，因此需要适中的阈值来剪除伪边。然而，当 τ 进一步提高至 0.9 时，性能有所下降，这说明过高的阈值会移除检测细微伪造内容所必需的关键语义线索。

结论： 在 *ForgeryNet* 均衡 子集上，我们观察到性能随着 τ 的增加而稳步提升，并在 $\tau = 0.7$ 处达到峰值（67.04%）。这表明，$\tau = 0.7$在过滤噪声与保留有意义的语义依赖关系之间实现了最佳权衡。因此，我们通过实证将 $\tau = 0.7$ 设为最终模型的默认配置。