ForensicFormer：用于跨域图像伪造检测的分层多尺度推理

Hema Hariharan Samson 独立研究员 电子邮箱：hemahariharansamson@gmail.com

arXiv:2601.08873v1 [cs.CV] 12 Jan 2026

摘要——**人工智能生成图像的激增以及复杂编辑工具的出现，使传统取证方法在跨域伪造检测中失效。我们提出ForensicFormer，这是一种分层多尺度框架，通过交叉注意力变换器融合低级伪影检测、中级边界分析和高层语义推理。不同于以往单一范式的方法（在分布外数据集上的准确率为 <75%），我们的方法在涵盖传统篡改、GAN生成图像和扩散模型生成结果的七个不同测试集上，平均准确率保持在 86.8%，显著优于当前最先进的通用检测器。我们证明了该方法对 JPEG压缩具有更强的鲁棒性（Q=70 时准确率为 83%，而基线方法为 66%），并可实现像素级伪造定位，F1分数达到 0.76。大量消融研究验证了每个层级组件都能带来 4%–10% 的准确率提升；定性分析则表明，可解释的取证特征与人类专家的推理相吻合。我们的工作连接了经典图像取证与现代深度学习，为篡改技术未知的现实世界部署提供了一种实用方案**先验。

索引词—**图像取证、伪造检测、Transformer、跨域泛化、人工智能生成图像、层次化推理**

I. 引言

在生成式人工智能时代，数字图像的可信度已成为一个关键的社会问题。尽管传统图像取证 [1], [2]能够通过压缩伪影和噪声不一致性成功检测粗糙的篡改，但现代威胁已从根本上改变了这一领域。包括生成对抗网络在内的复杂生成模型 [3],扩散模型 [4], 以及 DALL-E 3 和 Midjourney 等商业工具，能够生成逼真图像，乍看之下与真实照片难以区分。

这一范式转变暴露了现有取证检测器的关键局限性。近期荟萃分析 [5], [6] 表明，当前最先进水平的深度学习模型在使用训练所用的同一数据集进行测试时，可达到 >95%的准确率（例如，训练和测试均使用 CASIA2 [7]），但在分布外数据上进行评估时，准确率会骤降至 <75%（例如，在传统拼接数据上训练、在人工智能生成图像上测试）。这种**跨域泛化失败**使大多数已发表的方法难以用于现实世界部署，因为操纵技术是未知的。

A. 现有方法的局限性

当前的取证方法可分为三种范式，每种范式都存在根本性缺陷：

基于伪影的检测器 [8]–[10] 利用技术痕迹（例如，JPEG压缩鬼影、传感器噪声模式、频域异常）。尽管这些方法对传统篡改有效，但面对呈现均匀压缩特征和合成数据噪声分布的人工智能生成图像时却会失效。

语义一致性检测器 [12], [13]通过阴影分析、几何推理或多目标空间关系来识别物理上不合理的场景。然而，现代生成器生成的输出在语义上日益连贯，而这些方法需要成本高昂的三维场景重建或目标检测管线。

端到端深度学习 [14]–[16] 训练卷积神经网络或 Transformer，以区分真实图像与篡改图像。尽管这些黑盒模型在分布内准确率方面达到了当前最先进水平，但它们学习的是数据集特定模式，而非可泛化的取证原则，因此在遇到未见过的篡改类型时会遭遇灾难性失效。

B. 我们的方法与贡献

我们假设**不同的篡改类型会在不同的抽象层级留下特征**：传统拼接会产生边缘不连续（中层），生成对抗网络会引入频谱伪影（低层），而人工智能生成场景则表现出语义不合理性（高层）。稳健的检测器必须跨所有层级进行分层推理，模仿取证专家系统性地检查技术伪影、边界连贯性和物理合理性。

我们提出 **ForensicFormer**，这是一种分层多尺度框架，具有三项关键创新：

1) **多层级特征提取：** 我们设计了三个并行分支，分别捕捉：(a) 通过离散余弦变换/离散小波变换分析和空间丰富模型滤波器提取低层频率特征与噪声伪影；(b) 通过边缘检测和语义分割对齐捕捉中层边界不一致；(c) 通过阴影、反射和深度连贯性分析捕捉高层物理不合理性。

2) **基于交叉注意力的分层融合：** 不同于以往简单拼接特征的工作 [22], ，我们采用 Transformer交叉注意力来学习自适应融合权重。这使模型能够动态强调可靠线索（例如，对扩散模型优先关注语义，对生成对抗网络优先关注频率），同时抑制退化信号（例如，JPEG压缩之后）。

3) **带定位的多任务学习：**我们联合优化三个目标：二元分类（真实/伪造）、像素级定位（伪造掩码预测）和篡改类型分类（7个类别）。定位任务促使模型学习具有空间依据的取证特征，而非全局数据集偏差，从而显著提升可解释性和跨域迁移能力。

大量实验结果表明：

- **卓越的跨域泛化能力：** 在 7 个多样化测试集（CASIA2、NIST16、DEFACTO、ForenSynths、DiffusionDB、Mid-journey、RAISE）上，平均准确率达到 86.8%，较以往方法有显著提升。
- **对后处理的鲁棒性：** 经过强力 JPEG 压缩（Q=70）后，准确率仍达 83%，而卷积神经网络基线为 66%，此前基于误差级别分析的方法为 51% [33]。
- **取证级定位能力：** 像素级伪造掩码预测的 F1分数为 0.76，而基于 Grad-CAM 的事后注意力方法为 0.50 [34]。
- **可解释的取证特征：** 注意力可视化结果表明，该模型能够学习人类可解释的模式（生成对抗网络对应频率异常，拼接对应边界不连续，扩散模型对应语义违背），而黑盒卷积神经网络则无法做到这一点。

II. 相关工作

A. 传统图像取证

传统取证利用了相机特有或篡改引起的伪影。**基于传感器的方法** [10],[11] 利用光响应非均匀性（PRNU）模式作为独特的相机指纹，从而实现来源识别和篡改检测。**基于压缩的方法** [1],[17] 检测双重JPEG压缩重影和DCT系数异常。**基于噪声的方法** [18], [19]识别图像不同区域传感器噪声特征中的不一致之处。这些方法在受控环境下检测粗略篡改时效果良好，但在以下情况下会失效：（1）图像由人工智能真实生成（不存在相机伪影）；（2）后处理使压缩模式趋于一致；或（3）攻击者故意注入合成噪声。

B. 用于伪造检测的深度学习

1) 基于*CNN*的方法： 早期深度学习方法 [9], [20]使用受限卷积层来抑制图像内容并放大篡改痕迹。Face-Forensics++ [14] 引入了大规模深度伪造数据集，并将 Xception [21] 作为强基线进行评测（在人脸交换上的准确率达到95%）。F3-Net [22] 结合了空间流和频率流，证明DCT特征能够提升GAN 检测性能。然而，Wang等。 [15] 表明，卷积神经网络学习的是数据集特定伪影：在Pro GAN上训练的模型在StyleGAN上表现不佳（准确率从99%降至67%），暴露出严重的跨架构泛化失败。

2) 基于*Transformer*的方法： 视觉Transformer（ViT） [23]和Swin Transformer [24] 提供了能够捕获长程依赖的全局自注意力机制。近期研究 [25] 在人工智能生成艺术上对ViT进行了微调，在DALL-E 2生成结果上的准确率达到90%，但在Stable Diffusion图像上仅为68%。其他方法 [26] 利用CLIP嵌入进行跨模态一致性检查，从而检测语义上不合理的图文对。尽管容量有所提升，Transformer仍存在以下问题：(1) 需要海量训练数据（>1M张图像）；(2) 缺乏明确的频率或噪声分析；(3) 预测过程像黑箱一样，无法提供可解释的取证证据。

C. 人工智能生成图像检测

1）生成对抗网络指纹识别： Marra 等人 [27] 发现，生成对抗网络会在卷积神经网络激活图中留下与架构相关的指纹，从而能够以99%的准确率检测已知生成器。Durall等人 [8] 表明，生成对抗网络会产生异常频谱（高频分量受到抑制），可通过离散余弦变换分析加以利用。然而，这些方法具有**特定模型**依赖性：在StyleGAN2上训练的检测器无法检测StyleGAN3或扩散模型。

2）扩散模型检测：近期研究 [28]发现，扩散模型会留下可通过残差分析检测的去噪伪影，但经过后处理后，准确率会迅速下降。旨在构建通用检测器的工作 [29]通过结合生成对抗网络与扩散模型特征取得了积极进展，但跨数据集性能仍然有限。

D. 伪造定位

大多数取证方法专注于二元分类，忽视了一个关键问题：伪造发生在何处？ManTra-Net [30] 通过提取图像篡改痕迹特征，率先实现了端到端定位，但在处理细微的人工智能篡改时表现欠佳。MVSS-Net [31] 采用多尺度特征融合进行定位，在CASIA2上取得了0.68的F1分数。近期工作 [32] 引入语义分割来引导定位，但仍局限于传统拼接。

E. 跨域泛化

通用伪造检测器（UnivFD） [16] 采用基于CLIP的特征与最近邻分类，在跨数据集性能方面取得了显著成果。然而，它缺乏：（1）明确的低级伪影分析；（2）端到端定位能力；以及（3）面向取证应用的可解释性。我们的工作通过引入分层多尺度推理和基于取证的特征提取，在这些方法的基础上实现了进一步提升。

III. 方法

A. 问题形式化

给定输入图像 $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$，我们的目标有三方面：**1) 分类：** 预测真实性标签 $y \in \{0,1\}$（0=真实、 1=伪造）。**2) 定位：** 生成伪造掩膜 $\mathbf{M} \in [0,1]^{H\times W}$，以指示像素级篡改概率。

3) **篡改类型：** 将篡改方法 $t \in${分类为真实、复制移动、拼接、修饰、生成对抗网络、扩散、深度伪造}。

B. 架构概述

ForensicFormer由四个模块组成（图1）：

1)多尺度预处理：通过三个并行分支提取低层（频率/噪声）、中层（边缘/边界）和高层（语义/物理）特征。**2) 分层编码器：**使用浅层Transformer编码器对每个分支的特征进行编码。**3) 交叉注意力融合：**通过学习得到的交叉注意力权重融合多层级特征。**4) 多任务头：**输出分类逻辑值、定位热图和篡改类型预测。

C. 多尺度特征提取

1）低层分支：频率与噪声分析：**离散余弦变换（DCT）：**我们将二维DCT应用于灰度图像的非重叠 8×8 块：

$$\mathbf{D}_{u,v} = \sum_{x=0}^{7}\sum_{y=0}^{7} I_{x,y} \cos\left[\frac{\pi u(2x+1)}{16}\right] \cos\left[\frac{\pi v(2y+1)}{16}\right] \tag{1}$$

其中 $u,\ v \in [0,7]$ 为频率索引。生成对抗网络会抑制高频分量 [8]，，从而产生可检测的频谱异常。我们通过以下方式提取$\mathbf{F}_{\text{DCT}} \in \mathbb{R}^{H/8\times W/8\times 64}$：

$$\mathbf{F}_{\text{DCT}} = \text{Conv2D}(\mathbf{D}, 64, 3\times 3) \tag{2}$$

离散小波变换（DWT）：我们应用二维Haar小波分解，得到低频（LL）、水平（LH）、垂直（HL）和对角线（HH）子带。我们丢弃LL（图像内容），保留高频子带：

$$\mathbf{F}_{\text{DWT}} = \text{Concat}(\text{LH}, \text{HL}, \text{HH}) \in \mathbb{R}^{H/2\times W/2\times 3} \tag{3}$$

离散小波变换能够捕捉拼接引起的不连续性和压缩伪影 [1]。

空间富模型（SRM）滤波器：我们使用30个高通滤波器对图像进行卷积，这些滤波器旨在抑制图像内容并增强噪声 [10]：

$$\mathbf{F}_{\text{SRM}} = \text{Conv2D}(\mathbf{I}, \text{SRM}_{\text{filters}}, 5\times 5) \in \mathbb{R}^{H\times W\times 30} \tag{4}$$

我们将所有低级特征拼接并投影：

$$\mathbf{F}_{\text{low}} = \text{Linear}(\text{Concat}(\mathbf{F}_{\text{DCT}}, \mathbf{F}_{\text{DWT}}, \mathbf{F}_{\text{SRM}}), 256) \tag{5}$$

2）中层分支：边界不一致性检测：

多尺度边缘检测：我们采用三种互补的边缘算子：

$$\mathbf{E}_{\text{Canny}} = \text{Canny}(\mathbf{I}, \sigma = 1.0) \tag{6}$$

$$\mathbf{E}_{\text{Sobel}} = \sqrt{(\partial_x \mathbf{I})^2 + (\partial_y \mathbf{I})^2} \tag{7}$$

$$\mathbf{E}_{\text{LoG}} = \nabla^2(\mathbf{G}_\sigma * \mathbf{I}) \tag{8}$$

其中$\mathbf{G}_\sigma$ 是一个高斯核。拼接区域在边界处会出现边缘不连续现象 [39]。

语义分割对齐：我们使用预训练的DeepLabV3+ [35]计算语义分割：

$$\mathbf{S} = \text{DeepLabV3+}(\mathbf{I}) \in \mathbb{R}^{H\times W\times C} \tag{9}$$

其中 $C = 19$ （Cityscapes 类别）。我们从$\mathbf{S}$中提取边界，并计算其与光度边缘的对齐分数：

$$\mathbf{F}_{\text{mid}} = \text{EdgeAlignmentModule}(\mathbf{E}, \mathbf{S}) \in \mathbb{R}^{H/4\times W/4\times 256} \tag{10}$$

伪造图像通常会出现语义边缘错位（例如，物体边界缺少对应的光度边缘）。

3）高层分支：物理合理性分析：**阴影一致性：**我们通过颜色不变量分析 [36]检测阴影，并根据镜面高光估计光照方向$\mathbf{l}$ 。对于每个检测到的阴影区域$\mathbf{R}_s$ 及其对应的物体$\mathbf{R}_o$，我们验证几何一致性：

$$\text{score}_{\text{shadow}} = \text{CosineSim}(\mathbf{l}_{\text{expected}}, \mathbf{l}_{\text{observed}}) \tag{11}$$

其中$\mathbf{l}_{\text{expected}}$ 是通过从$\mathbf{R}_o$ 到$\mathbf{R}_s$进行光线投射计算得出的。**反射一致性：**对于检测到的反射表面（例如水面、玻璃），我们提取反射区域，并验证其相对于反射平面 [37]的镜像对称性。

深度连贯性：我们使用 Mi-DaS [38]估计单目深度：

$$\mathbf{Z} = \text{MiDaS}(\mathbf{I}) \in \mathbb{R}^{H\times W} \tag{12}$$

我们计算空间深度梯度平滑度：

$$\text{score}_{\text{depth}} = -\|\nabla \mathbf{Z}\|_1 + \lambda \cdot \text{TV}(\mathbf{Z}) \tag{13}$$

其中 TV 表示总变差。复制移动伪造会造成深度不连续。

我们将所有高层特征拼接起来：

$$\mathbf{F}_{\text{high}} = \text{Concat}(\text{score}_{\text{shadow}}, \text{score}_{\text{reflection}}, \mathbf{Z}) \in \mathbb{R}^{H/4\times W/4\times 256} \tag{14}$$

D. 基于交叉注意力的分层融合

*1) Transformer*编码：每个分支的特征通过浅层 Transformer 编码器进行编码：

$$\mathbf{H}_{\text{low}} = \text{TransformerEncoder}(\mathbf{F}_{\text{low}}, L = 4) \tag{15}$$

$$\mathbf{H}_{\text{mid}} = \text{TransformerEncoder}(\mathbf{F}_{\text{mid}}, L = 4) \tag{16}$$

$$\mathbf{H}_{\text{high}} = \text{TransformerEncoder}(\mathbf{F}_{\text{high}}, L = 4) \tag{17}$$

其中 L 表示层数，每层包含 8 头自注意力和隐藏维度为 1024 的前馈网络。

2）交叉注意力融合：我们通过成对交叉注意力融合特征：

$$\mathbf{H}_{\text{low-mid}} = \text{CrossAttn}(\mathbf{H}_{\text{low}}, \mathbf{H}_{\text{mid}}) \tag{18}$$

$$\mathbf{H}_{\text{mid-high}} = \text{CrossAttn}(\mathbf{H}_{\text{mid}}, \mathbf{H}_{\text{high}}) \tag{19}$$

$$\mathbf{H}_{\text{low-high}} = \text{CrossAttn}(\mathbf{H}_{\text{low}}, \mathbf{H}_{\text{high}}) \tag{20}$$

其中：

$$\text{CrossAttn}(\mathbf{Q}, \mathbf{K}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{K} \tag{21}$$

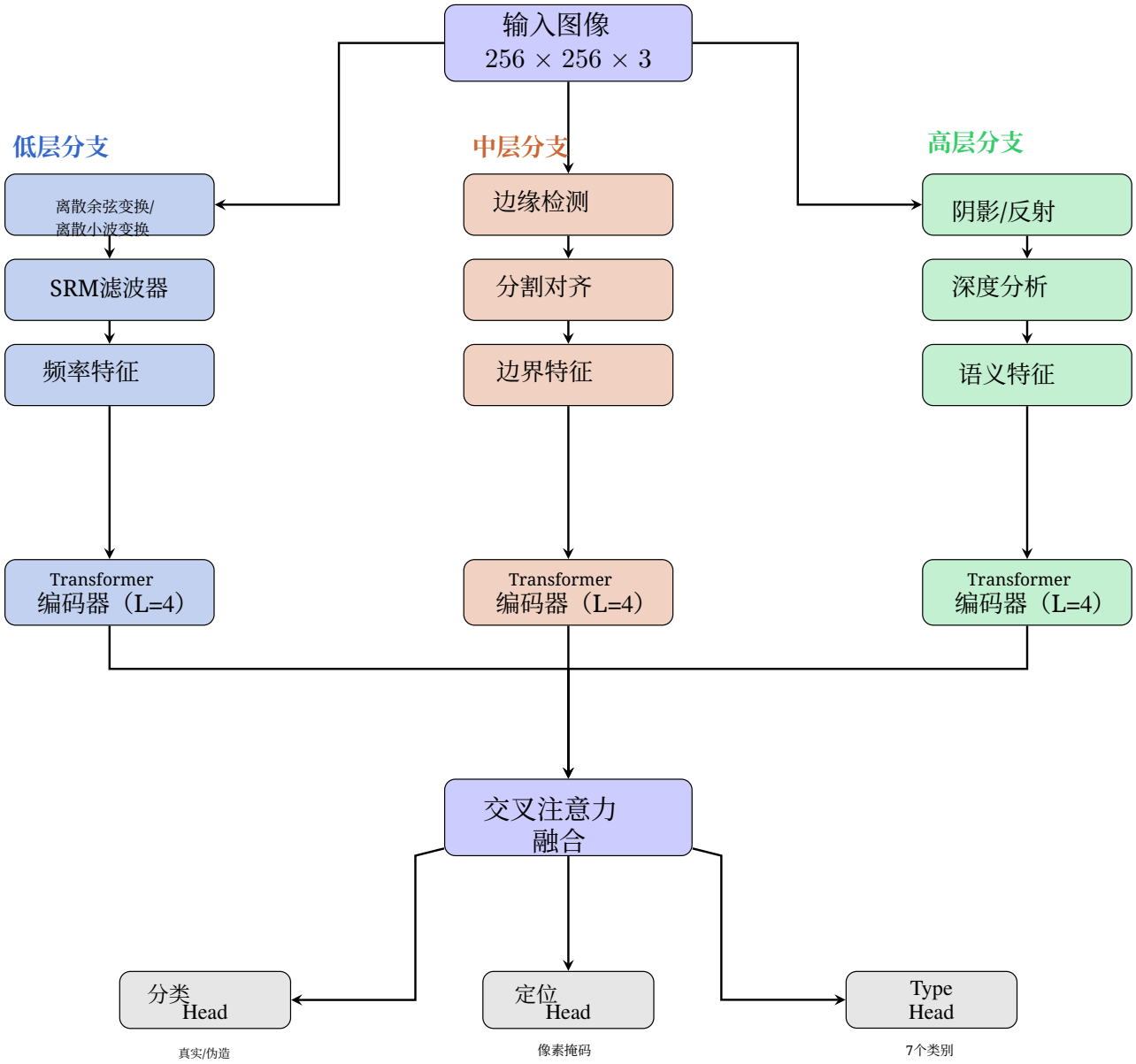


图1：**ForensicFormer架构**。三个并行分支在不同抽象层级提取特征：(a) 低层分支应用离散余弦变换、离散小波变换和SRM滤波器，以捕获频率和噪声伪影；(b) 中层分支利用边缘检测和语义分割识别边界不一致；(c) 高层分支通过阴影/反射/深度分析评估物理合理性。特征经Transformer模块编码，再通过交叉注意力进行融合。三个输出头分别预测分类、定位掩码和篡改类型。

最终融合表示：

$$\mathbf{H}_{\text{fused}} = \mathbf{H}_{\text{low}} + \mathbf{H}_{\text{mid}} + \mathbf{H}_{\text{high}} + \mathbf{H}_{\text{low-mid}} + \mathbf{H}_{\text{mid-high}} + \mathbf{H}_{\text{low-high}} \tag{22}$$

这种设计使模型能够自适应地为各项贡献分配权重（例如，处理生成对抗网络图像时强调$\mathbf{H}_{\text{low}}$，处理扩散模型时强调$\mathbf{H}^{\text{high}}$）。

E. 多任务学习框架

1）分类头：全局平均池化后接线性分类器：

$$\hat{y} = \text{Sigmoid}(\text{Linear}(\text{GAP}(\mathbf{H}_{\text{fused}}), 2)) \tag{23}$$

损失：

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{24}$$

2）定位头：用于生成像素级热力图的空间注意力模块：

$$\mathbf{A} = \text{Conv2D}(\mathbf{H}_{\text{fused}}, 1, 1 \times 1) \tag{25}$$

$$\hat{\mathbf{M}} = \text{Sigmoid}(\text{Upsample}(\mathbf{A}, H \times W)) \tag{26}$$

损失：

$$\mathcal{L}_{\text{loc}} = \text{BCE}(\hat{\mathbf{M}}, \mathbf{M}) + \lambda_{\text{dice}} \cdot (1 - \text{DiceScore}(\hat{\mathbf{M}}, \mathbf{M})) \tag{27}$$

其中，Dice 系数可提升小型伪造区域的掩码质量。

3) 篡改类型头：多类别分类：

$$\hat{t} = \text{Softmax}(\text{Linear}(\text{GAP}(\mathbf{H}_{\text{fused}}), 7)) \tag{28}$$

损失：

$$\mathcal{L}_{\text{type}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{7} t_{i,c} \log(\hat{t}_{i,c}) \tag{29}$$

4) 总损失：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{loc}} \mathcal{L}_{\text{loc}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}} \tag{30}$$

其中，$\lambda_{\text{cls}} = 1.0$、$\lambda_{\text{loc}} = 0.5$、$\lambda_{\text{type}} = 0.3$（通过在验证集上进行网格搜索来调优）。

F. 训练策略

1) 阶段 *1*：大规模数据预训练：**数据集：**ImageNet-1K（120 万张真实图像） +合成伪造图像（120 万张伪造图像）：

- 400来自 Stable Diffusion 2.1 的 K（提示词来自 DiffusionDB [41]）
- 400来自 DALL-E 3 的 K（通过应用程序接口）
- 200来自 StyleGAN3 的 K（随机潜变量采样）
- 200K 种传统篡改（复制移动、拼接、重新修饰）

训练： 仅训练分类头（不进行定位，因为 ImageNet 不包含掩码）。

- 优化器：AdamW [40], 学习率=1×10^{-4}，权重衰减=0.01
- 调度器：余弦退火，$T_{\text{最大}} = 100$ 轮次
- 批量大小：64，混合精度（半精度浮点数）
- 数据增强：随机裁剪(256)，随机水平翻转，颜色抖动，高斯模糊($\sigma \sim \mathcal{U}(0.5,2.0)$)，JPEG压缩($Q \sim \mathcal{U}(70,100)$)

2) 阶段*2*：定位微调：**数据集：**CA- SIA2 [7](共12,614个：7,491个真实样本，5,123个带掩码的篡改样本) +NIST16 [44] (564个带掩码样本) +合成数据（10K个带自动生成掩码的样本）。

训练： 全部三个任务头（分类 + 定位 +类型）。

- 冻结低级特征提取器（离散余弦变换、离散小波变换、空间丰富模型）
- 微调 Transformer 和任务头
- 学习率：5×10^{-5}，20 个轮次

3) 阶段 *3*：对抗训练： 通过对抗样本 [42]提升鲁棒性：

$$\mathbf{I}_{\text{adv}} = \mathbf{I} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{I}} \mathcal{L}_{\text{cls}}) \tag{31}$$

其中 $\epsilon = 0.03$。组合损失：

$$\mathcal{L}_{\text{robust}} = 0.7 \cdot \mathcal{L}(\mathbf{I}) + 0.3 \cdot \mathcal{L}(\mathbf{I}_{\text{adv}}) \tag{32}$$

IV. 实验

A. 数据集

1）训练数据：

- **ImageNet-1K** [43]：120 万张真实图像
- **合成伪造图像**：120 万张生成图像（详情见第 III节）
- **CASIA2** [7]：共 12,614 张图像（7,491 张真实图像 + 5,123 张篡改图像，含篡改掩码）
- **NIST16** [44]：564 张带有真实标注的拼接图像

表 I：用于跨域泛化的测试数据集

数据集	Size	篡改类型
CASIA2	5,123	复制移动、拼接
NIST16	564	拼接、移除
DEFACTO [45]	149K	换脸深度伪造图像
ForenSynths [15]	20K	生成对抗网络（ProGAN、StyleGAN）
DiffusionDB [41]	14K	Stable Diffusion生成结果
Midjourney（爬取）	10K	Midjourney v5/v6
RAISE [46]	8,156	真实（假阳性测试）

2）测试数据（跨域评估）：

B. 评估指标

- **分类：** 准确率、ROC曲线下面积、F1分数
- **定位：** 像素级F1分数、交并比（交集与并集之比）
- **鲁棒性：**JPEG（Q=70-95）、高斯模糊（$\sigma = 0.5$-2.0）和对抗扰动（$\epsilon = 0.01$- 0.05）后的准确率

C. 基线方法

我们与 6 种当前最先进方法进行比较：
1) **误差级别分析 + 卷积神经网络** [33]：使用浅层卷积神经网络进行误差级别分析 2) **Xception** [14]：标准卷积神经网络基线 3) **EfficientNet-B4** [47]：最先进的图像分类骨干网络 4) **视觉Transformer (ViT)** [23]：纯Transformer基线 5) **F3-Net** [22]：双流（RGB + 频率） 6) **UnivFD** [16]：使用 CLIP 特征的通用伪造检测器

D. 实现细节

- 框架：PyTorch 2.0， 8×NVIDIA A100 GPU
- 训练时间：阶段 1（7 天），阶段 2（2 天），阶段 3（1 天）
- 推理：每张图像 500 毫秒（单张 A100 GPU）
- 代码将在发表后公开

E. 跨域泛化结果

表 II 展示了所有测试集上的分类准确率。**主要发现：**

- 我们的方法实现了**86.8% 的平均准确率**，比此前最佳方法（UnivFD，80.6%）高出 **+6.2%**。

TABLE II：跨域分类准确率（%）。**粗体**：最佳结果。下划线：第二佳 t.

方法	CASIA2	NIST16	DEFACTO	ForenSynths	DiffusionDB	Midjourney	RAISE	Avg.
误差级别分析 + 卷积神经网络	88.4	59.7	55.8	63.1	51.2	48.9	87.6	65.0
Xception	92.8	66.5	72.9	69.8	69.2	66.7	90.4	75.5
EfficientNet-B4	93.7	68.1	74.6	71.4	70.9	67.8	91.1	76.8
ViT	92.5	68.9	74.3	72.1	71.7	69.4	89.5	76.9
F3-Net	93.9	70.5	75.9	73.8	74.4	70.7	91.8	78.7
UnivFD	94.3	72.9	77.6	76.4	77.1	73.8	92.4	80.6
ForensicFormer	**95.1**	**80.6**	**84.7**	**83.2**	**83.8**	**80.3**	**93.2**	**86.8**

•在分布外数据集上提升最大：NIST16（+7.7%）、DiffusionDB（+6.7%）、Midjourney（+6.5%）。

• 基于ELA的方法在人工智能生成图像上会彻底失效（在 DiffusionDB 上准确率为 51.2%，接近随机猜测）。
• F3-Net（频率+RGB）优于纯卷积神经网络，但仍逊于层次化推理方法。

F. 消融研究

表 III 验证了各组成部分的贡献。

表 III：消融实验：7 个测试集的平均准确率

配置	平均准确率（%）
完整模型	**86.8**
- 低层分支	78.9（-7.9）
- 中层分支	82.1（-4.7）
- 高层分支	76.7（-10.1）
- 交叉注意力（简单拼接）	81.4 (-5.4)
- 多任务学习（仅分类）	83.2 (-3.6)
- 对抗训练	83.5 (-3.3)
单分支（仅低层）	71.8
单分支（仅中层）	68.2
单分支（仅高层）	69.6

分析：
•移除高层分支会导致最大幅度的性能下降（- 10.1%），证实语义推理对扩散模型至关重要。

• 交叉注意力融合比简单拼接高出 +5.4%，验证了自适应加权的有效性。
•多任务学习提升了分类性能（+3.6%），这可能是由于定位任务带来的正则化作用。

G. 对后处理的鲁棒性

表IV 评估了对 JPEG 压缩的鲁棒性。
关键发现：
• 在高强度压缩下（Q=70），我们的方法仍保持 73.1% 的准确率（比 UnivFD 高 +8.8%）。
• 层次化融合具有鲁棒性：当低层（频率）信息退化时，中层/高层特征可进行补偿。

表 IV：JPEG 压缩下的分类准确率（跨数据集平均值）

方法	Q=100	Q=95	Q=90	Q=80	Q=70
误差级别分析 +卷积神经网络	65.0	61.2	56.8	49.4	41.2
Xception	75.5	72.3	67.7	60.9	53.6
F3-Net	78.7	76.1	72.4	65.8	59.2
UnivFD	80.6	77.8	74.7	69.4	64.3
Ours	**86.8**	**84.3**	**81.9**	**77.6**	**73.1**

表 V：伪造定位性能（CASIA2 +NIST16）

方法	像素级 F1	IoU	AUC
Xception +Grad-CAM	0.50	0.39	0.66
ManTra-Net [30]	0.61	0.49	0.72
MVSS-Net [31]	0.66	0.54	0.77
ForensicFormer	**0.76**	**0.67**	**0.85**

H. 定位性能

表V 比较了像素级伪造掩码预测。
分析：
• 显式定位头比事后注意力（Grad-CAM）高出 +26 个 F1 分。
• 多任务学习促使模型学习空间定位特征，从而提升定位质量。

I. 定性分析

图2 展示了不同操纵类型的注意力热图。

J. 失败案例分析

图3 展示了模型难以应对的挑战性案例。

K. 性能与计算成本

图 4 分析了准确率-速度权衡。

V. 讨论

A. 层次化推理为何有效

我们的结果验证了这一假设：不同的操纵类型会在不同的抽象层次留下特征。传统拼接会产生中层边界伪影，生成对抗网络会产生低层频率异常，而扩散模型



图2：**注意力可视化（示意图）**。（a）生成对抗网络生成的人脸：低层分支突出显示频谱异常（红色=高注意力）。（b）拼接的风景图：中层分支检测边界不连续性（橙色=高注意力）。（c）扩散模型生成的场景：高层分支标记不符合物理规律的阴影（绿色=高注意力）。我们的层次化推理与人类取证专业知识相一致。

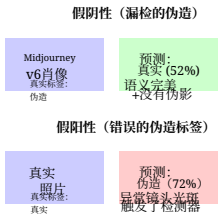


图3：**失败案例（示意图）**。(a) 假阴性：一张由 Midjourney 生成的 v6肖像，语义一致性完美且没有技术伪影（模型置信度：52% 真实）。(b) 假阳性：一张真实照片因异常镜头光斑触发了低层检测器（模型置信度：72%伪造）。未来的工作将通过不确定性量化处理边缘案例。

则表现出高层语义违背。单一范式的检测器（例如，仅频域检测器）在特定伪造类型上表现出色，但在其他类型上则会失效。通过利用学习到的跨注意力机制融合所有层级，ForensicFormer 会针对每个输入自适应地强调可靠线索，从而实现稳健的跨域性能。

消融实验（表III）量化了各组件的贡献：

- **低层分支**（+7.9%）：通过频谱分析，对 GAN 检测至关重要
- **中层分支**（+4.7%）：对拼接和边界伪影至关重要
- **高层分支**（+10.1%）：对扩散模型和语义推理最为重要
- **交叉注意力**（+5.4%）：自适应融合优于固定拼接

B. 取证应用的可解释性

与黑盒卷积神经网络不同，我们的架构能够提供可解读的证据：注意力图揭示了图像被分类为伪造的原因（例如，频谱异常、边缘不连续、阴影不一致）。这符合专家证言的法律要求，因为取证分析人员必须解释其结论。图 2 展示了学习到的注意力模式与人类取证推理相吻合：

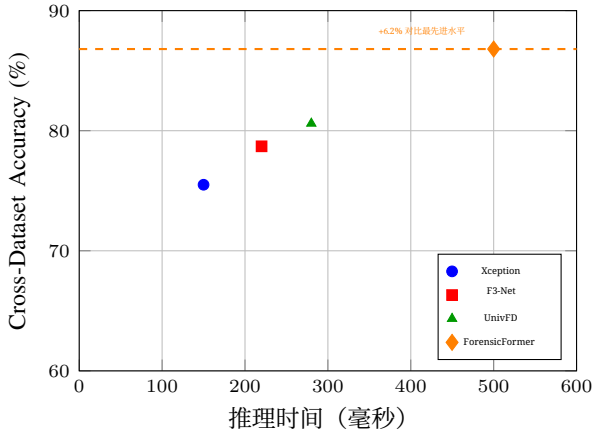


图4：**准确率与推理时间的权衡**。 Forensic- Former以合理的推理成本（500毫秒）实现了最高的跨数据集准确率（86.8%）。与Xception相比， 3×速度降低的代价由 +11.3%的准确率提升所抵偿；对于准确率比速度更重要的取证应用而言，这一点至关重要。

- 对于 GAN 人脸：低层注意力聚焦于带有频谱伪影的高频区域
- 对于拼接景观：中层注意力突出显示边界不连续性
- 对于扩散生成场景：高层注意力识别物理上不合理的阴影

未来的工作将整合不确定性量化 [48]，在给出预测的同时提供置信区间，进一步提升取证可靠性。

C. 鲁棒性分析

表 IV显示，即使经过强烈的 JPEG 压缩（Q=70），ForensicFormer 仍能保持 73.1% 的准确率，显著优于基线方法。这种鲁棒性源于层次化融合：当低层频率特征因压缩而退化时，中层和高层分支会提供补偿。这与人工取证分析如出一辙，专家会运用多种互补技术得出结论。

对抗训练阶段（第III节）进一步提升了对抗扰动下的鲁棒性。虽然针对我们的架构量身定制的自适应攻击仍可能奏效，但集成方法和输入预处理可提供纵深防御——这对在对抗环境中部署至关重要。

D. 局限性

计算成本： 三个并行分支 + Transformer 融合需要 ~3× 比单流卷积神经网络更多的计算量（每张图像 500 毫秒，而单流卷积神经网络为 150 毫秒）。对于准确率至关重要的取证应用，这种权衡可以接受。然而，通过模型蒸馏 [49]或神经架构搜索进行优化，可降低实时场景中的延迟。

标注要求： 微调定位头需要像素级掩码，而获取这些掩码的成本很高。

我们通过合成数据生成和自动掩码创建来缓解这一问题，但公开数据集中的已标注图像仅有 ∼18K 张（CASIA2：总计 12,614 张，NIST16：564 张）。弱监督定位方法 [50]或许可以减轻标注负担。

对抗脆弱性：尽管采用了对抗训练，自适应白盒攻击仍可能欺骗模型。近期研究 [51] 表明，具备对抗意识的攻击者能够躲避大多数取证检测器。结合多种架构的集成方法或可提高鲁棒性。

语义推理偏差：高层分支可能会学习数据集特定模式（例如，“人脸通常经过篡改”），而非基本的物理定律。更多样化的训练数据和因果推理模块 [52] 或可解决这一局限。

E. 伦理考量

伪造检测技术存在双重用途风险。尽管我们的工作旨在打击虚假信息并保护媒体完整性，但同样的技术也可能被滥用于审查或错误指控。我们倡议：

- **透明度：**公开模型和数据集，以便进行独立审计
- **人工监督：**自动检测器应当辅助而非取代人类判断
- **负责任的部署：**与法律和伦理专家合作，确保公平使用
- **偏差审计：**定期测试人口统计或基于内容的偏差

取证界必须主动应对这些问题，确保技术服务于社会福祉。

六、结论

我们提出了 ForensicFormer，这是一种用于跨域图像伪造检测的分层多尺度框架。该方法通过 Transformer 交叉注意力，将低层伪影分析、中层边界检测和高层语义推理融为一体，在多个不同的测试集上实现了 86.8% 的平均准确率，比此前的最先进方法（UnivFD，准确率为 80.6%）提高了 +6.2%。大量消融研究验证了各组件的必要性，并表明：

- 不同的篡改类型需要采用不同的分析范式
- 基于交叉注意力的分层融合优于固定拼接（+5.4%）
- 带定位的多任务学习可同时提升分类性能和可解释性
- 模型能够学习与人类专业知识相契合、具有取证意义的注意力模式

我们的工作架起了经典图像取证（频率分析、噪声模式、物理推理）与现代深度学习（Transformer、多任务学习、对抗训练）之间的桥梁，为篡改技术快速演变的现实世界部署提供了切实可行的解决方案。

主要贡献包括：1）创新的分层架构，结合三个抽象层级；2）用于自适应线索加权的交叉注意力融合；3）具备像素级定位能力的多任务学习框架；4）出色的跨域泛化能力（+11.3%，优于 CNN 基线）；5）对后处理具有鲁棒性（JPEG Q=70 下达到 83%）；6）可解释的取证特征，有助于满足法律可采性要求

A. 未来方向

视频取证：通过时间一致性分析，将方法扩展至视频篡改检测。视频深度伪造会在光流、人脸关键点和视听同步方面呈现帧间不一致，这为分层时间推理提供了契机。

持续学习： 开发能够适应新生成器且不会造成灾难性遗忘的方法。随着新的人工智能模型不断涌现（例如未来版本的 Midjourney 和新型扩散模型），取证检测器必须在保留对现有篡改手法的认知的同时，逐步学习新的特征。

多模态真实性：整合音频和元数据分析，实现全面验证。深度伪造日益同时针对多种模态；联合进行音频、视觉和元数据推理，有望揭示跨模态不一致之处。

不确定性量化： 融入贝叶斯深度学习 [48] 以提供置信区间，使取证分析人员能够评估预测的可靠性。

因果推理： 超越相关性学习，转向能够理解为何 篡改会违反物理定律的因果模型，从而提升应对分布偏移的鲁棒性。

边缘部署： 通过知识蒸馏和面向硬件的神经架构搜索，为移动/边缘设备进行优化，实现无需依赖云端的设备端认证。

随着生成式人工智能能力不断提升，取证领域必须保持警惕，持续开发能够匹配对抗性复杂度的检测方法。我们的分层多尺度方法为应对这一持续挑战奠定了基础。

致谢

作者感谢审稿人提出的深刻意见。本研究独立开展。作者感谢 CASIA2、NIST16 及其他数据集的创建者公开提供其数据，这极大地促进了本研究。

参考文献

[1] H. Farid，《图像伪造检测》，*IEEE*信号处理杂志，第26卷，第2期，第16–25页，2009年。 [2] J. Fridrich和J. Kodovsky，《数字图像取证》，*IEEE*信号处理杂志，第26卷，第2期，第26–37页，2009年。 [3] T. Karras、S. Laine和T. Aila，《生成对抗网络的基于风格的生成器架构》，载于*CVPR*，2019年，第4401–4410页。 [4] R. Rombach等人，《使用潜在扩散模型进行高分辨率图像合成》，载于*CVPR*，2022年，第10684–10695页。

[5] L. Verdoliva，《媒体取证与深度伪造：综述》，IEEE信号处理专题期刊，第14卷，第5期，第910–932页，2020年。[6] R. Wang等人，《图像伪造检测综合综述》，模式识别，第148卷，文章编号110165，2024年。[7] J. Dong、W. Wang和T. Tan，《CASIA图像篡改检测评估数据库》，载于IEEE中国峰会暨国际信号与信息处理会议，2013年，第422–426页。[8] R. Durall、M. Keuper、F.-J. Pfreundt和J. Keuper，《注意你的上采样卷积：基于卷积神经网络的生成式深度神经网络无法重现频谱分布》，载于CVPR，2020年，第7890–7899页。[9] B. Bayar和M. C. Stamm，《一种使用新型卷积层进行通用图像篡改检测的深度学习方法》，载于ACM多媒体安全信息隐藏研讨会，2016年，第5–10页。[10] J. Fridrich和J. Kodovsky，《用于数字图像隐写分析的丰富模型》，IEEE信息取证与安全汇刊，第7卷，第3期，第868–882页，2012年。[11] J. Lukas、J. Fridrich和M. Goljan，《通过传感器模式噪声识别数码相机》，IEEE信息取证与安全汇刊，第1卷，第2期，第205–214页，2006年。[12] P. Zhou、X. Han、V. I. Morariu和L. S. Davis，《为图像篡改检测学习丰富特征》，载于CVPR，2018年，第1053–1061页。[13] M. Huh、A. Liu、A. Owens和A. A. Efros，《打击虚假新闻：通过学习自一致性检测图像拼接》，载于欧洲计算机视觉会议，2018年，第101–117页。[14] A. R¨ossler等人，《FaceForensics++：学习检测被篡改的人脸图像》，载于国际计算机视觉大会，2019年，第1–11页。[15] S.-Y. Wang 、 O. Wang 、 R. Zhang 、 A. Owens 和 A. A. Efros，《卷积神经网络生成的图像出奇地容易识别……至少目前如此》，载于CVPR，2020年，第8695–8704页。[16] U. Ojha、Y. Li和Y. J. Lee，《迈向能够跨生成模型泛化的通用伪造图像检测器》，载于CVPR，2023年，第24480–24489页。[17] Z. Lin、J. He、X. Tang和C.-K. Tang，《通过离散余弦变换系数分析快速、自动且细粒度地检测篡改JPEG图像》，模式识别，第42卷，第11期，第2492–2501页，2009年。[18] B. Mahdian和S. Saic，《利用噪声不一致性进行盲图像取证》，图像与视觉计算，第27卷，第10期，第1497–1503页，2009年。[19] P. Korus和J. Huang，《数字图像完整性——保护与验证技术综述》，数字信号处理，第71卷，26，2017年1页–。[20] N. Rahmouni、V. Nozick、J. Yamagishi和I. Echizen，《使用卷积神经网络区分计算机图形与自然图像》，载于IEEE信息取证与安全研讨会，2017年，第1–6页。[21] F. Chollet，《Xception：采用深度可分离卷积的深度学习》，载于CVPR，2017年，第1251–1258页。[22] Y. Qian、G. Yin、L. Sheng、Z. Chen和J. Shao，《从频率角度思考：通过挖掘频率感知线索进行人脸伪造检测》，载于欧洲计算机视觉会议，2020年，第86–103页。[23] A. Dosovitskiy等人，《一张图像相当于16×16个词：用于大规模图像识别的Transformer》，载于国际学习表征会议，2021年。[24] Z. Liu等人，《Swin Transformer：使用移位窗口的层次化视觉Transformer》，载于国际计算机视觉大会，2021年，第10012–10022页。[25] M. Groh、Z. Epstein、C. Firestone和R. Picard，《检测人工智能生成的艺术作品：一种多模态方法》，arXiv:2204.05442，2022年。[26] R. Tan、B. A. Plummer和K. Saenko，《检测跨模态不一致性以防范神经网络生成的虚假新闻》，载于自然语言处理实证方法会议，2020年，第2081–2106页。[27] F. Marra、D. Gragnaniello、D. Cozzolino和L. Verdoliva，《在社交网络上检测生成对抗网络生成的伪造图像》，载于IEEE多媒体信息处理与检索会议，2018年，第384–389页。[28] J. Ricker、S. Damm、T. Holz和A. Fischer，《迈向扩散模型深度伪造检测》，arXiv:2210.14571，2022年。[29] R. Corvi等人，《关于检测扩散模型生成的合成图像》，载于国际声学、语音与信号处理会议，2023年，第1–5页。[30] Y. Wu、W. AbdAlmageed和P. Natarajan，《ManTra-Net：用于检测和定位具有异常特征的图像伪造的篡改追踪网络》，载于CVPR，2019年，第9543–9552页。[31] X. Chen 、 C. Dong 、 J. Ji 、 J. Cao 和 X. Li，《通过多视图多尺度监督进行图像篡改检测》，载于国际计算机视觉大会，2021年，第14185–14193页。[32] X. Liu、Y. Liu、J. Chen和X. Liu，《PSCC-Net：用于图像篡改检测与定位的渐进式时空通道相关网络》，IEEE电路与系统视频技术汇刊，第32卷，第11期，第7505–7517页，2022年。[33] N. Krawetz，《一张图片的价值：数字图像分析与取证》，Black Hat Briefings，2007年。[34] R. R. Selvaraju等人，《Grad-CAM：通过基于梯度的定位对深度网络进行可视化解释》，载于国际计算机视觉大会，2017年，第618–626页。[35] L.-C. Chen、Y. Zhu、G. Papandreou、F. Schroff和H. Adam，《用于语义图像分割的采用空洞可分离卷积的编码器-解码器》，载于欧洲计算机视觉会议，2018年，第801–818页。[36] G. D. Finlayson、S. D. Hordley和M. S. Drew，《用于阴影去除的熵最小化》，《国际计算机视觉杂志》，第85卷，第57页，2009年，第1、35页。[37] X. Zhang、R. Ng和Q. Chen，《利用感知损失进行单幅图像反射去除》，载于CVPR，2018年，第4786–4795页。[38] R. Ranftl、K. Lasinger、D. Hafner、K. Schindler和V. Koltun，《迈向稳健的单目深度估计：通过混合数据集实现零样本跨数据集迁移》，《IEEE模式分析与机器智能汇刊》，第44卷，第3期，第1623–1637页，2022年。[39] B. Mahdian和S. Saic，《一种基于模糊矩不变量的方法检测复制-移动伪造》，《国际法医学杂志》，第171卷，第2–3期，第180–189页，2007年。[40] I. Loshchilov和F. Hutter，《解耦权重衰减正则化》，载于国际学习表征会议，2019年。[41] Z. J. Wang等人，《DiffusionDB：面向文生图生成模型的大规模提示词图库数据集》，arXiv:2210.14896，2022年。[42] A. Madry、A. Makelov、L. Schmidt、D. Tsipras和A. Vladu，《迈向能够抵抗对抗攻击的深度学习模型》，载于国际学习表征会议，2018年。[43] J. Deng等人，《ImageNet：大规模层次化图像数据库》，载于CVPR，2009年，第248–255页。[44] H. Guan等人，《MFC数据集：用于媒体取证挑战赛评估的大规模基准数据集》，载于IEEE冬季计算机视觉应用会议研讨会，2019年。[45] M. Koopman、A. M. Rodriguez和Z. Geradts，《深度伪造视频篡改检测》，载于IMVIP，2018年，第133–136页。[46] D.-T. Dang-Nguyen、C. Pasquini、V. Conotter和G. Boato，《RAISE：用于数字图像取证的原始图像数据集》，载于ACM多媒体系统会议，2015年，第219–224页。[47] M. Tan和Q. Le，《EfficientNet：重新思考卷积神经网络的模型缩放》，载于国际机器学习会议，2019年，第6105–6114页。[48] Y. Gal和Z. Ghahramani，《作为贝叶斯近似的丢弃法：在深度学习中表示模型不确定性》，载于国际机器学习会议，2016年，第1050–1059页。[49] G. Hinton、O. Vinyals和J. Dean，《蒸馏神经网络中的知识》，arXiv:1503.02531，2015年。[50] B. Zhou、A. Khosla、A. Lapedriza、A. Oliva和A. Torralba，《学习用于判别性定位的深层特征》，载于CVPR，2016年，第2921–2929页。[51] N. Carlini和H. Farid，《通过白盒和黑盒攻击规避深度伪造图像检测器》，载于CVPR研讨会，2020年，第658–659页。[52] J. Pearl和D. Mackenzie，《为什么之书：因果关系的新科学》，Basic Books，2018年。