

基于双感知网络的伪造引导学习策略用于 Deepfake 跨域检测

Lixin Jia¹, Zhiqing Guo^{1,2,*}, Gaobo Yang³, Liejun Wang^{1,2}, and Kegin Li⁴

¹School of Computer Science and Technology, Xinjiang University, Urumqi, China

²Xinjiang Multimodal Intelligent Processing and Information Security Engineering Technology Research Center, Urumqi, China

³College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

⁴Department of Computer Science, State University of New York, New Paltz, New York, USA

*Corresponding author: guozhiqing@xju.edu.cn

ABSTRACT

深度伪造技术的出现引入了一系列社会问题，引起了广泛关注。当前的深度伪造检测方法在特定数据集上表现良好，但在应用于未知伪造技术的数据集时表现不佳。此外，随着新兴和传统伪造技术之间的差距不断扩大，依赖于常见伪造痕迹的跨域检测方法变得越来越无效。这种情况突显出开发具有强泛化能力的深度伪造检测技术以应对快速迭代的伪造技术的紧迫性。为了解决这些挑战，我们提出了一种伪造引导学习（FGL）策略，旨在使检测网络能够持续适应未知的伪造技术。具体而言，FGL 策略捕捉已知和未知伪造技术之间的差异信息，使模型能够实时动态调整其学习过程。为了进一步提高对伪造痕迹的感知能力，我们设计了一个双重感知网络（DPNet），以捕捉伪造痕迹之间的差异和关系。在频率通道中，网络动态感知并提取各种伪造技术中的区分特征，建立基本的检测线索。然后将这些特征与空间特征结合并投射到嵌入空间。此外，图卷积用于感知整个特征空间的关系，促进对伪造痕迹相关性的更全面理解。大量实验证明，我们的方法在不同场景中具有良好的泛化性，并有效处理未知伪造挑战，为深度伪造检测提供了强大的支持。我们的代码可在 <https://github.com/vpsg-research/FGL> 获得。

Keywords Deepfake detection, forgery guided learning, frequency perception, graph convolution

1 介绍

在过去的十年中，深度学习的广泛应用促进了 deepfake 技术的迅速发展。各种面部伪造技术相继出现，例如 NVAE [?]、LDM [?]。利用这些伪造技术，非专业人员也能轻松生成伪造的图像和视频。因此，大量伪造内容被上传到主流社交媒体平台，带来了显著的社会风险。随着对生成对抗网络（GANs）[?] 的研究不断深入，现有的 deepfake 技术 [?] 一直在不断改进。因此，由 deepfake 生成的面部伪造图像越来越逼真，肉眼难以辨识。如果被恶意行为者滥用，deepfake 技术可能导致金融欺诈、社会不稳定甚至政治危机。

目前，已经提出了多种深度伪造检测方法 [?, ?] 来解决深度伪造技术带来的问题。这些方法在公共数据集上如 FaceForensics++ [?] 和 Celeb-DF [?] 取得了很高的检测性能。然而，检测模型主要依赖于这些数据集中存在的特定伪造痕迹来识别伪造内容。因此，它们在实际场景中的效果有限，特别是在面对未知伪造技术时。在实际环境中，新的伪造技术不断出现，并在互联网快速传播，这对有效检测这些技术产生的内容构成了巨大挑战。

为了克服这一挑战，广泛的研究 [?, ?, ?] 已经集中在通过捕捉不同技术间常见的伪造痕迹来检测深度伪造内容。例如，一些研究发现深度伪造视频通常在特定区域表现出微妙的伪影或不自然的纹理 [?], [?]，例如面部、眼睛或嘴巴的边缘。这些伪影可以作为区分伪造内容的重要特征。受元学习 [?] 和小样本学习 [?] 理论的启发，一些深度伪造检测算法尝试推广到未知伪造域 [?], [?]。尽管这些方法在跨域检测中展示了一些效果并显示出检测未知伪造技术的潜力，但它们仍然表现出一定的局限性。(1) 随着伪造技术的不断发展，已知和未知技术之间的差距正在扩大。最新的伪造模型可以在没有明显伪影的情况下创造高分辨率的伪造内容。因此，依赖识别常见伪造痕迹的跨域检测方法将逐渐变得不那么有效。(2) 在特征空间中不同域的伪造样本的分布高度不一致，这显著加剧了源域和目标域之间特征分布的偏差。此外，这种偏差阻碍了模型捕捉未知伪造模式的能力。(3) 基于元学习的方法旨在逐步适应新伪造样本的独特特征，但这通常导致先验知识的灾难性遗忘。此外，静态更新策略难以有效应对不断演变的伪造技术。

在这项工作中，我们通过引入伪造引导学习（FGL）策略来解决上述限制，该策略增强了检测网络对未知伪造技术的适应能力，并提高了跨域检测性能。如图 1 底部所示，网络通过 FGL 策略捕获已知和未知伪造技术之间的不同伪造信息。通过分析这些差异，模型可以实时动态调整其更新方向和幅度。该策略有助于减少

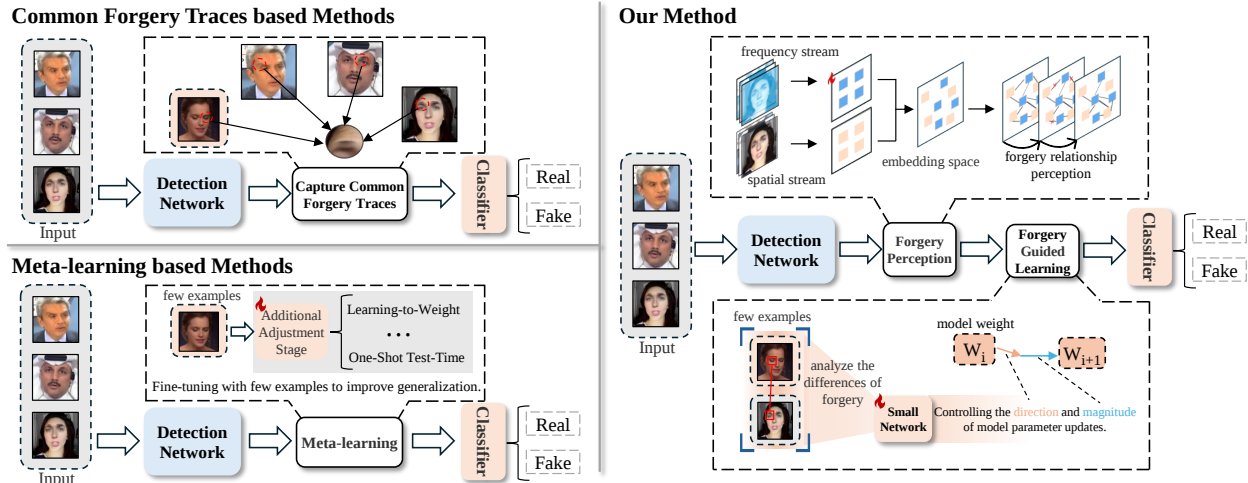


Figure 1: 与主流方法对比进行 deepfake 跨域检测。在图中，背景为橙色的脸部图像代表那些由未知伪造技术生成的图像。如图底部所示，我们的方法分析伪造技术之间的差异并逐步更新模型权重。此外，频率流动态地捕捉各种伪造特征，然后将其与空间流整合以获得更具区分性的嵌入空间。

域不变伪造特征的遗忘，并增强模型检测未知伪造模式的能力。由于已知和未知伪造样本在特征分布上的差异，我们设计了一种新颖的双感知网络（DPNet），以感知伪造特征之间的差异和关系。在其帮助下，FGL 能够有效利用从已知伪造技术中提取的具有代表性和可迁移的共享伪造特征，从而学习更通用的特征表示。

本研究的具体创新和贡献如下：

- 我们提出了一种新颖的 FGL 策略，该策略通过捕捉已知和未知伪造之间的特征差异来自适应地调整模型参数。基于先验知识和这些特征差异，该策略对当前的损失和梯度提供动态反馈，使模型能够逐步优化。
- 一种新颖的频域感知机制（FPM）被结合到 DPNet 的频率流中。FPM 通过动态路由机制动态地捕捉频域中的伪造特征，并将其与空间流特征整合，以构建一个更加具有辨别力的嵌入空间。
- 为了减轻特征冗余并增强伪造特征之间的交互，我们将图卷积理论整合到 DPNet 中。具体来说，自适应伪造关系感知（AFRP）模块在嵌入空间中学习特征表示，捕捉伪造和非伪造特征之间的关系，同时动态调整它们。
- 我们在五个数据集上进行了广泛的实验，模拟了各种真实世界场景，以全面验证我们方法的有效性。结果表明，所提出的方法凭借少量实例可以显著提高对未知伪造技术的检测，突出了其重要的应用价值。

本文的其余部分组织如下。第 ?? 节回顾了关于深度伪造检测技术的相关工作。第 2 节介绍了我们提出的方法。第 ?? 节报告了不同实验场景下的评估结果。第 ?? 节总结了本文，并讨论其局限性和未来方向。

近年来，许多面部伪造检测方法被开发出来，以减轻与深度伪造技术滥用相关的风险。早期的检测方法主要集中在识别伪造图像中的局部伪影和不一致性。例如，Face X-ray 方法突出显示伪造面部区域与背景图像之间的混合差异，这些差异作为检测 Manipulation 的显著指示符。Li 等人展示了深度伪造视频通常会表现出独特的视觉伪影，可以通过卷积神经网络（CNN）有效捕获，以区分面部图像的真实性。这些发现显著推进了深度伪造检测技术的发展。

除了分析空间域中的伪造痕迹外，研究人员还在频域中识别出假图像和真实图像之间的显著差异。F3Net [?] 将频域信息整合到 CNN 中，使得能够有效检测伪造人脸图像中的细微操纵痕迹。然而，最近的研究 [?] 显示出不同的伪造技术产生不同的频域伪影。如图 2 所示，我们使用快速傅里叶变换（FFT）对各种 deepfake 数据集进行频谱分析，揭示出数据集之间的细微差异。受到这些观察的启发，我们提出了频域感知机制（FPM）。与依赖于静态频域伪造线索的现有方法不同，我们的 FPM 渐进捕捉频域中的样本特定伪造线索。通过将动态频率与空间信息结合，我们构建了一个更具判别力的嵌入空间，从而提高伪造检测的准确性和鲁棒性。

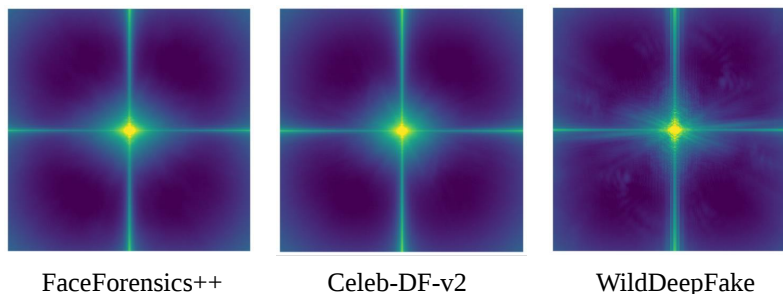


Figure 2: 各种 Deepfake 数据集的频率分析。每个 FFT 频谱图代表从相应数据集中随机选择的 2,000 张图像的平均结果。

1.1 基于图的深度伪造检测

一个新兴的趋势是将图神经网络（GNN）应用于计算机视觉领域。Vision GNN [?] 提出了将图像表示为图结构的新方法。通过将图像划分为作为节点的块并结合图卷积操作来构建邻接关系，解决了传统 GNN 中的过度平滑问题。同时，GNN 为提高深度伪造检测的泛化能力提供了一种新方法。杨等人 [?] 将深度伪造检测视为图分类问题，其中每个面部区域对应一个顶点，提供了一种新的视角。同时，他们发现图中大量冗余关系信息会阻碍图的表达能力，因此提出了一种掩码关系学习方法来减少冗余信息。

同样，伪造特征信息的冗余也会阻碍检测网络的判别精度。为了解决这个问题，我们将自适应伪造关系感知 (AFRP) 引入检测网络，并与 FPM 一起形成双感知网络。AFRP 将嵌入空间中的特征视为节点，并利用邻接矩阵感知并动态调整它们的关系。与参考文献 [?] 中的掩码关系学习方法主要减少冗余不同，AFRP 专注于增强特征表示。这使得模型能够更好地关注语义相关和信息丰富的特征。

随着越来越多的深度伪造数据集和伪造技术的出现，实现有效的跨域检测已成为一个关键的研究挑战。因此，许多研究 [?], [?] 提出了不同的方法来提高检测模型的泛化能力。例如，赵等人 [?] 提出了一种多注意力深度伪造检测网络，以捕捉真实与伪造图像之间的细微差异。DFGaze [?] 分析面部视频帧中的凝视特征，以更有效地识别真实和伪造面孔之间的线索。董等人 [?] 提出了一种不依赖身份信息的深度伪造检测模型，通过减少对图像身份信息的学习来增强泛化能力。尽管这些方法在处理常见伪造技术时表现良好，但当面对缺乏明显共享特征的未知伪造时，其效果是有限的。

其他研究提出了新的方法来应对新兴的伪造技术。通过结合元学习技术，Sun 等人 [?] 设计了一个 Learning-to-Weight 框架。他们认为不同的面孔在多个领域中对模型的贡献不同，这可能在特定领域导致模型偏差。在元测试阶段，损失函数被用来指导模型更新，选择最适合目标领域的模型。Ni 等人 [?] 提出了一种一致表示学习方法，通过不同的增强来捕获多样化的表示，然后应用正则化以增强伪造检测的一致性。在另一项基于元学习的研究中 [?]，通过在应用于测试样本之前更新模型来提高模型的适应性。这是通过构建特定于测试样本的辅助任务来实现的。

然而，这些检测方法采取的静态更新策略在增量学习过程中显示出明显的遗忘现象，最终导致模型的泛化能力下降。本文提出了一种新的解决方案，称为伪造引导学习 (FGL) 策略。FGL 通过学习源域和目标域伪造样本之间的特征差异，逐步引导模型更新。此外，我们根据当前模型的参数和梯度自适应地调整每个更新步骤的方向和幅度，从而提高模型的适应性和泛化能力。

2 提出的方法

在本节中，我们介绍所提出的方法。整体框架如图 3 所示。我们将原始数据集视为训练集 ($\mathcal{T}$)，将少量示例视为支持集 ($\mathcal{S}$)。我们的框架分为两个阶段。在第一阶段，双重感知网络 (DPNet) 在 $\mathcal{T}$ 上进行训练，以学习基本的伪造模式。频域感知机制 (FPM) 逐步提取频域中多个层次的伪造线索，捕捉不同尺度的伪造特征。随后，自适应伪造关系感知 (AFRP) 模块分析这些伪造特征之间的关系，并根据其变化动态调整。在第二阶段，我们提出了一种伪造引导学习 (FGL) 策略，该策略利用少量未知伪造技术的支持集来引导模型的学习过程。以下部分详细介绍了我们的方法。

2.1 伪造引导学习策略

为了增强检测模型对未知伪造技术的适应性，我们引入了伪造引导学习 (FGL) 策略，该策略使模型能够有效地在不同的伪造模式间泛化。整个位图策略在图 3 (b) 中进行了说明。我们将通过双重感知网络 (DPNet)

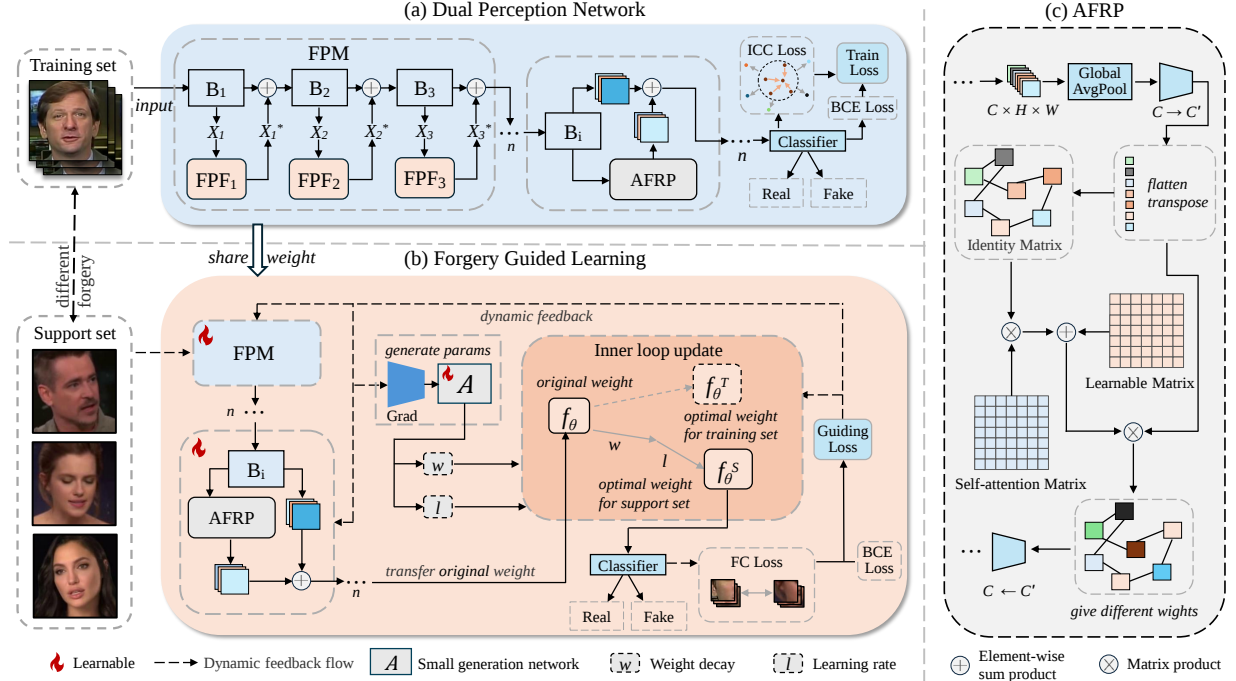


Figure 3: 我们所提出方法的总体架构。图 (a) 展示了双感知网络 (DPNet)。训练集被输入模型中，并通过频域感知机制 (FPM) 引入动态频域信息。然后，关键信息通过多个自适应伪造关系感知 (AFRP) 加以区分。最终，ICC 损失辅助 BCE 损失指导模型学习。图 (b) 代表了伪造引导学习 (FGL) 策略。在此阶段，该策略被提出用于指导检测模型提高其在未知伪造技术上的性能。FGL 通过学习支持集和原始数据集之间的差异来调整模型参数。在此阶段的损失由 FC 损失和 BCE 损失共同构成。图 (c) 提供了 AFRP 的详细说明。

学习的基本伪造模式作为先验知识，记为 f_θ ，它代表整个模型的权重参数。随后，使用 FGL 策略逐步更新 f_θ ，并使用相应的损失函数 L_S^T 进行评估，以评估其对未知伪造技术的泛化能力。

传统检测方法通常旨在寻找一个最佳权重，以在不同的造假技术中实现良好的泛化。但是，这样获得的权重通常是训练集的最佳权重 f_θ^T 。现有研究显示，设计一种更好的快速适应方法可以在学习少量样本时提高对任务的适应性 [?]。受到这一观点的启发，我们的方法着重通过 FGL 策略适应未知的造假领域，以获得未知造假技术的最佳权重 f_θ^S 。为了防止模型对少量样本过拟合 S ，我们引入了一个 l_2 正则化项 $\frac{\lambda}{2} \|\theta\|_2$ 。我们的 FGL 策略中的基本参数更新在方程 (1) 中提出。在

$$\begin{aligned} \theta_{i+1} &= \theta_i - l(\nabla_\theta L_S^T(f_{\theta_i}) + \lambda \theta_i) \\ &= w \theta_i - l \nabla_\theta L_S^T(f_{\theta_i}), \end{aligned} \quad (1)$$

中， i 表示步骤索引， f_θ 表示来自 DPNet 的模型权重参数， θ 表示更新后的权重参数。这两个控制变量， l 和 w ，控制模型参数更新的方向和幅度。本质上，学习率和权重更新量由从当前样本计算的损失和梯度决定，标记为 l 。另一方面， w 控制正则化的程度，具体而言，决定现有知识应经历权重衰减的幅度。

仅仅依靠这两个控制变量不足以应对所有可能的伪造检测情况，因此我们在更新模型参数时需要更高的灵活性。因此，这两个控制变量可以用可调节变量 l_i 和 w_i 替换，它们与 $\nabla_\theta L_S^T(f_{\theta_i})$ 和 θ_i 具有相同的维度。最后，我们的 FGL 策略方程变为

$$\theta_{i+1} = w_i \odot \theta_i - l_i \odot \nabla_\theta L_S^T(f_{\theta_i}), \quad (2)$$

，其中 $\odot$ 表示按元素相乘。FGL 策略的目的是逐步引导模型适应未知的伪造技术。为了更精确地引导模型的学习过程，我们在步骤 i 基于特定的学习状态 t_i 生成控制变量。在所提出的框架中，控制变量 l_i 和 w_i 由一个小型生成网络 $\mathcal{A}$ 生成，如下：

$$(l_i, w_i) = \mathcal{A}(t_i). \quad (3)$$

如图 3 (b) 所示，小生成网络 $\mathcal{A}$ 在每次内循环更新之前生成特定的控制变量。我们使用两层 MLP 构建了 $\mathcal{A}$ ，在层之间包含一个 ReLU 激活函数。 $\mathcal{A}$ 接受一个向量 t_i 作为输入，该向量的大小是基础模型 DPNet 中层数的两倍。输出 l_i 和 w_i 按层生成，然后扩展以匹配各自参数 θ_i 的维度。这些控制变量随后被用作学习率和权重衰减因子，以调整内循环每一步期间参数更新的方向和幅度。

在引导检测模型学习的过程中，模型的更新依赖于每一步的当前学习状态。我们定义学习状态 t_i 由当前模型权重 θ_i 和对应的梯度 $\delta_i = \nabla_{\theta} L_S^T(f_{\theta_i})$ 组成。为了计算权重和梯度的层次平均，我们计算基础模型所有层的平均值如下：

$$\bar{\theta}_i = \frac{1}{n} \sum_{l=1}^n \theta_i^l, \quad \bar{\delta}_i = \frac{1}{n} \sum_{l=1}^n \delta_i^l, \quad (4)$$

其中 l 表示模型中的层数， $\bar{\theta}_i$ 和 $\bar{\delta}_i$ 分别表示平均权重和梯度。这导致以下学习状态：

$$t_i = [\bar{\theta}_i, \bar{\delta}_i]. \quad (5)$$

学习状态 t_i 不仅捕捉了当前模型的权重和梯度分布，还通过其动态变化反映了引导学习过程的方向和速度。一般来说，在 FGL 的初始阶段，模型参数更新较多，梯度也较大，最终趋于稳定。

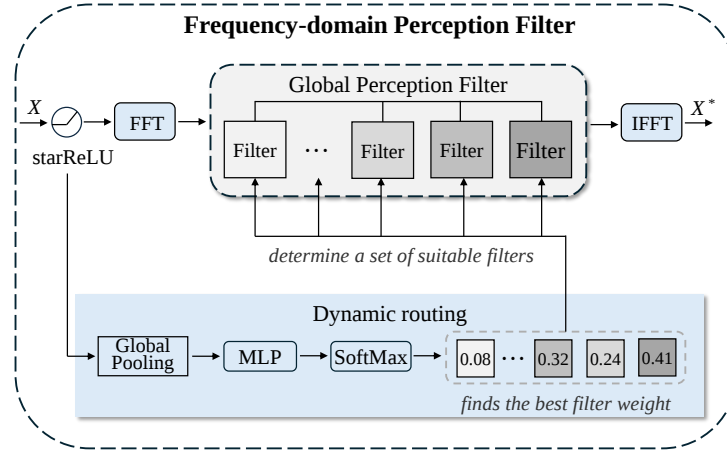


Figure 4: FPF 的详细结构。我们使用动态路由机制来找到最适合当前输入的滤波器权重。然后，这些权重被分配给不同的滤波器，以动态确定一个全局感知滤波器。此外，我们使用 starReLU 作为我们的激活函数。

2.2 双重感知网络

2.2.1 频域感知机制

由于不同的伪造技术将在频域中表现出不同的特性，我们设计了一种频域感知机制 (FPM)，如图 3 (a) 左侧所示。受到动态滤波器 [?] 的启发，我们使用一组频域感知滤波器 (FPF) 动态调整提取的多层级信息，引入来自不同层级的频域特征。如图 4 所示，FPF 使用多个滤波器动态确定全局感知滤波器。FPF 的思想可以简单描述如下：

$$X^* = \mathcal{F}^{-1} \left(\sum_{k=1}^K w_k(X) \cdot (W_f^{(k)} \odot \mathcal{F}(\mathcal{R}(X))) \right), \quad (6)$$

其中 X 表示输入特征张量， X^* 是携带动态频域信息的最终输出特征， $\odot$ 表示逐元素相乘。我们用 $\mathcal{F}$ 和 $\mathcal{F}^{-1}$ 表示快速傅里叶变换及其逆变换。 $W_f^{(k)}$ 表示第 k 个频域滤波器的权重 $w_k(X)$ 由动态路由机制 [?] 生成。 $\mathcal{R}(X)$ 是输入的激活函数。在这里，我们使用了由 Yu 提出的 starReLU 激活函数，其定义如下：

$$\mathcal{R}(X) = s \cdot \text{ReLU}(X)^2 + b. \quad (7)$$

通过使用一个可学习的比例 s 和一个偏差 b ，激活函数可以在训练过程中动态调整，增强了方法的灵活性。

动态路由机制通过输入特征在频域中自适应地找到最佳的滤波器权重。根据输入数据的特征，我们在不同滤波器之间分配权重，以动态选择最适合当前输入的全局感知滤波器。具体而言，输入的全局特征首先由全局池化 $G(X)$ 提取，以描述整个输入特征的总体信息。然后，全局特征通过一个轻量级的多层感知器 (MLP) 映射到动态权重空间。最后，生成权重 $w_k(X)$ 的过程如下：

$$w_k(X) = \frac{\exp(\phi_k \cdot \text{MLP}(G(X)))}{\sum_{j=1}^K \exp(\phi_j \cdot \text{MLP}(G(X)))}, \quad (8)$$

ϕ_k 是一个可学习的标量，通过训练进行优化，用于衡量每个滤波器的重要性。其中， exp 和归一化项确保生成的权重 $w_k(X)$ 满足概率分布 $\sum_{k=1}^K w_k(X) = 1$ 。

通过在权重生成过程中引入 ϕ_k ，过程不仅依赖于输入特征，还能够自适应地调整每个滤波器的初始重要性。动态生成的权重 $w_k(X)$ 用于对滤波器 $W_f^{(k)}$ 进行加权，然后与频域特征 X_f 进行逐元素相乘，结果 X_w 中包含动态频域信息，这有助于检测模型识别频域中的独特伪造痕迹。然后这些特征与空间域中的伪造特征集成并在嵌入空间中投射。

在引入动态频域信息后，嵌入空间中伪造特征的表示得到了进一步丰富。然而，某些冗余或无关的信息可能对模型的性能产生负面影响。此外，自适应图卷积已经被证明能够关注语义相关和信息丰富的特征。为了解决这个问题，我们引入了自适应伪造关系感知 (AFRP) 模块，并将其集成到双感知网络 (DPNet) 中，以促进显著信息的提取，如图 3 (c) 所示。AFRP 的核心概念是在嵌入空间中特征建模为图的顶点，并利用图卷积理论建立它们之间的依赖关系，从而实现伪造特征更具结构性和差异化的表示。

首先，我们将处理输入的特征图。具体来说，输入特征图 $x \in \mathbb{R}^{B \times C \times H \times W}$ 经过全局平均池化 $G(x)$ ，然后被展平并转置成一个特征顶点矩阵。特征图的处理表达如下：

$$\begin{cases} V' = \text{Conv}_r(G(X)), & V' \in \mathbb{R}^{B \times C' \times 1 \times 1} \\ V = \text{Transpose}(\text{Flatten}(V')), & V \in \mathbb{R}^{B \times 1 \times C'} \end{cases} \quad (9)$$

其中 $C' = \frac{C}{r}$ ，我们使用 Conv_r 来降低特征图的维度。为了降低计算复杂度，我们引入了一个瓶颈结构，将特征维度从 C 降低到 $\frac{C}{r}$ ，然后再恢复。这一操作也提高了特征之间相互作用的效率。

为了描述特征顶点之间的依赖关系，AFRP 设计了一个由三个组成部分构成的邻接矩阵。首先，单位矩阵 $M_I = I$ 表示特征顶点之间的自连接关系。接着，使用自注意力矩阵 M_{sa} 来强调每个特征顶点的权重。最后，可以学习的全局邻接矩阵 M_l 捕捉到任意两个特征顶点之间的全局关系。最终的邻接矩阵 M 通过以下方式获得：

$$\begin{cases} M = M_I \odot M_{sa} + M_l, \\ M_I = I, & M_I \in \mathbb{R}^{C' \times C'} \\ M_{sa} = T(S(\text{Conv}(V))), & M_{sa} \in \mathbb{R}^{C' \times C'} \\ M_l = \text{Parameter}(1 \times 10^{-6}), & M_l \in \mathbb{R}^{C' \times C'} \end{cases} \quad (10)$$

其中 S 代表 Softmax 函数， T 代表将结果整理为对角矩阵。参数指的是在模型梯度下降过程中更新的可学习参数。我们将参数初始化为 1×10^{-6} ，以保证学习的稳定性。

最后，生成的邻接矩阵 M 用于图卷积操作，以调整特征的重要性。并通过 Conv_{ir} 将特征维度恢复到 $X' \in \mathbb{R}^{B \times C \times H \times W}$ 。整个过程可以表示为：

$$\begin{cases} X' = X \cdot \sigma(\text{Conv}_{ir}(\text{ReLU}(\text{Conv}(V \cdot A))))), \\ X' \in \mathbb{R}^{B \times C \times H \times W}, \end{cases} \quad (11)$$

，其中 σ 表示使用 Sigmoid 激活函数对输出权重进行归一化。通过 AFRP 学习这些特征的权重关系，模型在全局范围内增强有用特征，同时抑制无关特征。这种全局特征处理方法不仅提高了模型对未知伪造技术的适应性，还提升了其在未知数据集上的性能。

由于 DPNet 和 FGL 策略旨在实现不同的目标，因此为每个部分采用独立的损失函数以确保最佳学习和性能。对于 DPNet 的损失函数，我们使用 Sun 等人提出的类内紧凑 (ICC) 损失来辅助二元交叉熵 (BCE) 损失进行训练。ICC 损失的基本思想是聚集正样本并使所有负样本远离真实中心 C_{real} 。通过采用这种损失，我们可以促进模型探索更多的区分特征并提高模型的泛化能力，这与我们的研究理念相符。ICC 损失定义为：

$$\begin{cases} L_{icc} = L_{positive} - L_{negative}, \\ L_{positive} = \frac{1}{|O^{real}|} \sum_{j=1}^{|O^{real}|} (o_j^{real} - C_{real})^2, \\ L_{negative} = \frac{1}{|O^{fake}|} \sum_{j=1}^{|O^{fake}|} (o_j^{fake} - C_{real})^2, \end{cases} \quad (12)$$

其中 O^{real} 和 O^{fake} 分别代表正样本和负样本集。然后使用 λ 将加权的 ICC 损失与 BCE 损失结合以获得最终损失：

$$L_{\mathcal{T}} = L_{bce} + \lambda L_{icc}. \quad (13)$$

]

对于 FGL 的损失函数，我们从 InfoNCE 损失 [?] 汲取灵感，并将特征对比 (FC) 损失整合到我们的目标函数中，以增强特征的区分能力和学习效率。首先，从模型中提取的特征 $X \in \mathbb{R}^{B \times C \times H \times W}$ 被展平为 $F \in \mathbb{R}^{B \times D}$ ，并计算它们之间的相似度矩阵：

$$S_{i,j} = \frac{X_i \cdot X_j}{\|X_i\| \|X_j\| \cdot \tau}, \quad S_{i,j} \in \mathbb{R}^{B \times B} \quad (14)$$

，其中 τ 是控制相似度分布平滑性的温度参数。最后，给出 FC 损失的公式为：

$$L_{fc} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(S_{i,i})}{\sum_{j=1}^B \exp(S_{i,j})}. \quad (15)$$

在 FGL 中，参数更新依赖于模型计算的任务梯度。FC 损失提供了来自特征相似性的额外梯度信息，这与分类损失不同。通过引入对比损失，增强了梯度的多样性。我们还计算了来自源域和目标域的损失。此阶段的总体损失为：

$$L_S^T = L_{bce}^T + \mu L_{bce}^S + \nu L_{fc}. \quad (16)$$

为了评估我们提出方法的有效性，我们模拟了真实场景，并在五个公共的 deepfake 检测数据集上进行了实验，重点关注以下两个实际场景。(1) 在第一个场景中，我们使用广泛采用的 FaceForensics++ (FF++) 数据集作为训练集。通过评估其在不同数据集上的性能，我们评估了该方法应用于未知数据集时的泛化能力。(2) 在第二个场景中，我们在 FF++ 数据集上进行实验，其中四种伪造技术中的三种用于训练，而剩余的未知伪造方法用于测试。该设置使我们能够评估所提方法处理先前未知伪造技术的有效性。

2.3 数据集

FaceForensics++ (FF++) [?] 被广泛认为是该领域中使用最为普遍的数据集。它提供了多个优点，包括大量的数据、多样化的伪造技术以及按压缩质量分组的视频样本。该数据集包含 1000 个原始真实视频，以及使用四种不同伪造技术生成的 1000 个对应伪造视频：Deepfakes (DF) [?]、Face2Face (F2F) [?]、FaceSwap (FS) [?] 和 NeuralTextures (NT) [?]。

Celeb-DF-v2 (CDF) [?] 数据集是一个高质量的 deepfake 数据集，专为名人面孔而设计，提供出色的视觉真实感。它包含 590 个真实视频和 5639 个对应的伪造视频。

DeeperForensics-1.0 (DFR) [?] 数据集包含 50,000 个真实视频和 10,000 个伪造视频。它不仅提供了大量数据，还包含多样的姿势、光照条件和表情，使其更具代表性，贴近真实场景。

WildDeepFake (WDF) [?] 数据集包含 3,805 个真实面部视频和 3,509 个伪造面部视频。真实视频和伪造视频均来自互联网，提供了多样化的内容，增加了检测的挑战性。

DeepFake Detection Challenge (DFDC) [?] 构建了大规模的 DFDC 数据集，该数据集包含 23,654 个由 3,426 名被雇用的演员在各种环境中拍摄的真实视频。此外，还包括 104,500 个使用更先进的深度伪造技术制作的伪造视频。

对于所有数据集，我们遵循官方文档来划分数据集。我们从每个视频中采样 20 帧，构建我们的实验数据集。我们使用 MTCNN 来裁剪面部区域并将其调整为 224×224 。所有实验都是在使用单个 NVIDIA RTX4090D 的 PyTorch 框架下进行的。模型使用从 ImageNet-1K 的预训练参数初始化。我们使用 Adam 优化器，并将参数 α 和 β 分别设定为 0.999 和 0.99。初始学习率 L_r 被设置为 2×10^{-4} ，并且每 5 个 epoch 衰减 0.5 倍。我们训练每个检测模型 20 个 epoch，批量大小为 32。在训练阶段，我们还应用了数据增强技术以提高数据多样性。

2.3.1 评价指标

我们遵循当前主流评估策略，使用准确率 (ACC) 和接收者操作特性曲线下面积 (AUC) 作为实验的评估指标。此外，我们通过测量参数的数量 (Param.) 和浮点运算 (FLOPs) 来评估模型的复杂度。

我们将我们提出的方法与最新和代表性的方法进行比较，包括 XceptionNet [?]、F3-Net [?]、LTW [?]、MAT [?]、RECCE [?]、CORE [?]、CADDMM [?]、IFFD [?]、UMFC [?]、SFICov [?]、DFGaze [?]。为了确保公平比较，我们使用作者提供的代码重现了结果。此外，实验中使用的数据集和实验设置与我们的方法保持一致。所有实验结果都使用图像级指标进行评估，这对于检测方法来说更具挑战性。

首先，我们在 FF++ 数据集上评估我们的方法，如表 1 所示。最佳结果用粗体显示，第二优秀的结果用下划线表示。在众多优秀的方法中，我们的方法和 RECCE [?] 同时在 FF++ 数据集中达到了最佳性能。与基于元学习的 LTW 方法 [?] 相比，我们的方法在数据集中实现了更优的性能，因为双重感知网络 (DPNet) 能够实

Table 1: FF++ 上的数据集内评估结果

Method	Venue	FF++(c23)		Params.	FLOPS
		ACC(%)	AUC(%)		
Xception[?]	ICCV 2019	75.60	88.63	20.81M	6.00G
F3-Net[?]	ECCV 2020	93.15	96.74	21.17M	8.49G
LTW[?]	AAAI 2021	94.31	97.80	20.37M	4.20G
MAT[?]	CVPR 2021	93.72	97.26	47.88M	5.18G
RECCE[?]	CVPR 2022	<u>95.26</u>	98.22	23.81M	6.17G
CORE[?]	CVPR 2022	91.75	96.99	20.81M	4.60G
CADDM[?]	CVPR 2023	93.96	97.45	-	4.83G
IFFD[?]	TIP 2023	94.06	97.39	20.81M	4.59G
UMFC[?]	AAAI 2024	91.61	95.48	54.01M	9.16G
SFICnv[?]	TIFS 2024	87.53	93.57	13.95M	3.09G
DFGaze[?]	TIFS 2024	89.44	90.90	8.54M	3.32G
FGL	Ours	95.38	<u>98.11</u>	20.19M	4.80G

现更全面的信息感知。通常，检测器在数据集上的表现与其从数据集中提取有用的伪造特征信息的能力相关。大多数深度伪造检测方法已经在数据集中实现了良好的性能。RECCE [?] 方法通过其重构-分类学习进一步实现了 98.22 % 的 AUC。我们方法的优势在于它可以自适应地判断伪造特征信息的重要性，并最终实现最高的 ACC 性能。F3-Net [?] 同样使用频域伪造信息。相比之下，我们的方法利用频域感知机制 (FPM) 捕捉不同的伪造模式，实现了频域信息的更有效利用，并显著增强了模型性能。其他检测方法，如 DFGaze [?] 和 SFICnv [?]，更多地强调检测模型的泛化性能，因此在数据集上的表现相对一般。对于 DFGaze [?]，我们保持了与其他模型相同的实验设置，训练了 20 个周期，而不是原论文中使用的 100 个周期。

Table 2: 未知伪造数据集的评估结果

Method	Venue	CDF		DFR		WDF		DFDC		Average
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	AUC(%)
XceptionNet[?]	ICCV 2019	64.77	68.52	51.21	70.50	61.46	68.52	63.61	69.63	69.29
F3-Net[?]	ECCV 2020	<u>72.99</u>	<u>74.03</u>	65.09	84.13	63.55	72.13	<u>65.20</u>	67.76	<u>74.51</u>
LTW[?]	AAAI 2021	63.40	64.10	60.90	84.90	59.10	63.40	63.10	69.00	70.35
MAT[?]	CVPR 2021	70.79	73.87	65.59	83.78	<u>65.03</u>	<u>73.33</u>	61.96	65.88	74.22
RECCE[?]	CVPR 2022	67.09	65.96	50.40	87.13	61.74	69.08	62.53	66.99	72.29
CORE[?]	CVPR 2022	68.78	68.67	71.41	<u>90.10</u>	62.75	68.04	61.45	67.36	73.54
CADDM[?]	CVPR 2023	68.75	66.83	67.48	87.05	62.23	70.71	64.90	68.72	73.33
IFFD[?]	TIP 2023	61.53	61.88	57.18	83.60	63.39	69.80	60.05	62.92	69.55
UMFC[?]	AAAI 2024	69.63	70.68	56.72	83.44	64.62	69.81	63.71	68.64	73.14
SFICnv[?]	TIFS 2024	68.76	70.36	56.67	78.82	63.01	69.92	64.83	67.96	71.77
DFGaze[?]	TIFS 2024	58.99	72.52	53.09	78.09	61.34	67.42	57.09	<u>73.54</u>	72.89
FGL	Ours	74.52	77.14	<u>69.31</u>	93.75	65.94	74.06	68.69	74.14	79.77

随着 deepfake 视频数据集数量的增加及其之间的显著差异，仅在特定数据集内取得良好表现是不够的。deepfake 面部检测的主要挑战是如何在未见的数据集上实现有效的泛化。我们提出的方法旨在实现在跨数据集场景中的有效检测。为了验证该方法的有效性，我们在 FF++ (c23) 上训练模型，并对不同的数据集进行性能评估，包括 CDF、DFR、WDF 和 DFDC。

表格 2 总结了跨数据集评估的结果。在此实验中，应用了与对比方法相同的设置，并且没有使用额外的支持集。我们提出的方法在其他方法中表现出明显的优势，在四个数据集上的平均结果显著更好。方法如 RECCE [?] 和 LTW [?] 在 FF++ 数据集上表现良好。然而，它们过于关注特定的伪造技术，导致在跨数据集评估中表现不佳。我们提出的方法在数据集内评估和跨数据集评估的结果之间实现了平衡。值得注意的是，F3-Net [?] 和 MAT [?] 方法在 CDF 和 WDF 数据集上表现良好，但在其他两个数据集上表现较差。对于我们的方法来

Table 3: 对未知伪造技术的评估结果

Method	Venue	TEST-DF		TEST-F2F		TEST-FS		TEST-NT		Average
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)	
F3-Net[?]	ECCV 2020	82.31	89.50	63.85	70.00	49.68	54.04	57.85	63.29	69.21
MAT[?]	CVPR 2021	80.13	88.32	62.51	69.47	51.88	53.98	56.61	66.19	69.49
LTW[?]	AAAI 2021	77.00	87.42	63.86	74.22	51.22	55.27	53.68	64.32	70.31
LTW (5-shot)	AAAI 2021	78.23	87.37	<u>66.98</u>	72.69	52.43	<u>56.34</u>	55.94	67.41	70.95
CORE[?]	CVPR2022	75.13	85.08	66.53	<u>75.86</u>	49.92	52.41	58.65	66.50	69.96
CADDMM[?]	CVPR 2023	70.39	84.92	59.56	68.25	50.85	52.90	55.36	64.26	67.58
IFFD[?]	TIP 2023	82.26	90.38	66.85	73.75	50.45	53.98	<u>59.87</u>	66.91	71.26
SFiConv[?]	TIFS 2024	75.08	90.04	63.53	71.56	50.72	55.81	58.70	63.72	70.38
DFGaze[?]	TIFS 2024	<u>82.52</u>	91.13	63.73	70.94	51.57	55.66	57.44	65.82	70.89
FGL	Ours	83.13	<u>91.83</u>	63.50	75.45	<u>52.57</u>	55.35	54.57	<u>67.47</u>	<u>72.53</u>
FGL (5-shot)	Ours	81.25	91.97	71.27	78.29	55.25	59.15	64.16	70.53	74.64

说，没有倾向于任何特定的数据集，因为它引导模型参数达到在所有数据集上表现良好的值。然而，我们的方法在 DFDC 数据集上未能取得最佳结果。我们推测，DFDC 数据集中的伪造特征与 FF++ 有显著不同，这阻止了我们的方法找到适用于该数据集的良好泛化参数。

在现实世界中，不断涌现新的伪造技术，这对深度伪造检测任务构成了重大挑战。一个好的检测方法必须在面对未知的伪造技术时保持良好的性能。在本节中，我们通过在未知的伪造技术上评估模型来模拟现实场景。具体来说，我们通过在每个实验中从 FF++ 数据集中排除四种伪造方法之一来模拟现实世界的场景，以代表一种新兴的伪造技术。

表 3 展示了对不同未知伪造技术进行评估的实验结果。我们可以得出结论，大多数方法在面对未知伪造方法时并不能很好地检测到。当检测像 DF 这样的伪造方法时，模型表现良好，这些方法较旧且展示明显的伪造痕迹。然而，当处理 FS 和 NT 等较新的方法时，其性能显著下降，因为这些伪造痕迹较为细微。更重要的是，不同的深伪技术之间存在很大的差异。我们的方法使用少量数据来学习这种差异并提高模型检测未知伪造技术的能力。在实验中，我们展示了我们方法的两个版本：一个是不使用支持集的版本，另一个是使用支持集的版本 (5-shot)。在 FGL 中，我们仅使用我们的 DPNet 取得了良好的性能。在 FGL (5-shot) 中，我们应用了我们提出的 FGL 策略，在 5-shot 支持数据集中学习未知的伪造技术，从而达到最佳性能。

由于开源元学习方法的有限性，我们在实验中选择了 LTW [?] 进行比较。LTW [?] 在元测试阶段使用虚拟更新，但模型的学习能力并未显著增强，导致在检测未知伪造技术时的表现一般。相比之下，FGL 能有效引导模型适应未知伪造技术。在 F2F、FS 和 NT 数据集上的测试中，它取得了显著改善，其中在 NT 数据集上的准确率提高了近 10 %。因此，我们得出结论：在真实场景中，我们的方法表现良好，并能有效应对新兴的伪造技术。

2.4 消融研究

在本节中，我们进行了全面的消融实验，以彻底评估我们提出的方法。具体而言，我们的消融研究包括以下组件：1) 评估所提出方法的整体有效性；2) 演示伪造引导学习 (FGL) 策略的有效性。

Table 4: 评估不同组件的有效性

Method	FF++CDF	FF++DFR	FF+WDF	FF+DFDC	Average
	AUC(%)	AUC(%)	AUC(%)	AUC(%)	AUC(%)
Baseline	67.64	90.65	68.26	67.94	73.62
w/ FPM	68.6	92.87	69.18	73.76	76.10
w/ AFRP	73.64	92.21	71.53	71.53	77.22
DPNet (FPM+AFRP)	77.14	93.75	74.06	74.14	79.77

Table 5: FGL 策略有效性的评估

Method	TEST-F2F		TEST-NT	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)
w/o FGL	63.50	75.45	54.57	67.47
w/ FGL (1-shot)	70.05	78.56	61.00	72.98
w/ FGL (5-shot)	71.27	78.29	64.16	70.53

为了研究我们提出的方法中每个组件的贡献，我们进行了系列消融实验。我们的方法主要依赖于伪造引导学习 (FGL) 策略和双感知网络 (DPNet) 来进行伪造检测。然而，由于 FGL 策略作为一个额外的学习过程，并且本质上增强了模型的泛化能力，因此本节着重评估 DPNet 的影响。具体来说，我们消除了 DPNet 中的频域感知机制 (FPM) 和自适应伪造关系感知 (AFRP)，以评估它们各自的贡献。实验在 FF++ 上进行了训练，并在其他四个数据集上进行了测试，4 中的表格展示了 AUC 结果。FPM 将频域信息集成到我们的检测框架中，使得能够动态提取不同频率分量相关特征。随着 AFRP 模块的引入，模型的性能显著提升，展示了我们的自适应图卷积在强调重要特征信息中的有效性。因此，所提出的框架在设计的方法中实现了稳健的性能。

2.4.1 FGL 策略的有效性

FGL 策略的核心是引导模型使用少量伪造数据适应未知伪造方法。我们在三种配置下评估了我们的策略：没有 FGL、FGL (1-shot) 和 FGL (5-shot)。实验在 FF++ 数据集中不同的伪造方法上进行，详细结果如表 5 所示。我们选择了 F2F 和 NT 这两种更难检测的未知伪造技术作为测试集。实验显示，我们的 FGL 策略可以更加有效地检测未知伪造技术，仅需 1-shot 数据。具体而言，FGL (1-shot) 配置在 F2F 上达到了 70.05 % 的平均检测准确率，在 NT 上达到了 61.00 %，分别比基线提高了 6.55 % 和 6.43 %。1-shot 和 5-shot 配置之间的性能差距很小，显示了我们少样本学习方法的效率。这些结果证明我们的 FGL 策略显著增强了模型对新颖伪造技术的适应能力。

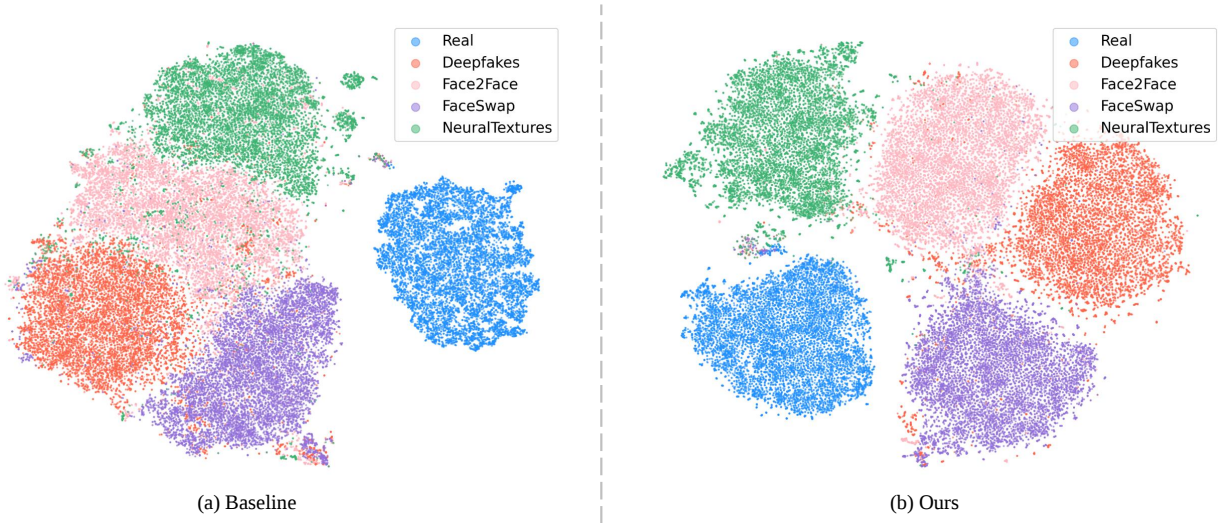


Figure 5: 基线我们的 FGL 的 T-SNE 特征分布可视化。我们的方法不仅可以区分真实图像和伪造图像，还可以学习不同伪造技术之间的差异。

在本节中，我们展示了输入特征分布和显著图的可视化，以证明我们 FGL 检测方法的有效性。

2.4.2 T-SNE 特征嵌入可视化

我们使用 t-SNE [?] 可视化 Baseline 和我们提出的方法的输入特征分布。FF++ [?] 数据集上的可视化结果如图 5 所示。由于不同的伪造方法表现出不同的操作伪迹，我们的方法不仅需要区分真实图像，还需要区分各种伪造技术。这些差异引导模型学习并适应未知的伪造技术。可视化结果清楚地显示，与基线相比，我们的方法实现了更好的类间分离。特别是对于像 DeepFakes [?] 和 Face2Face [?] 这样具有挑战性的伪造方法，我们的方法显示出与真实图像和其他操作类型明显分开的特征分布模式。这种改进的特征辨别能力直接有助于

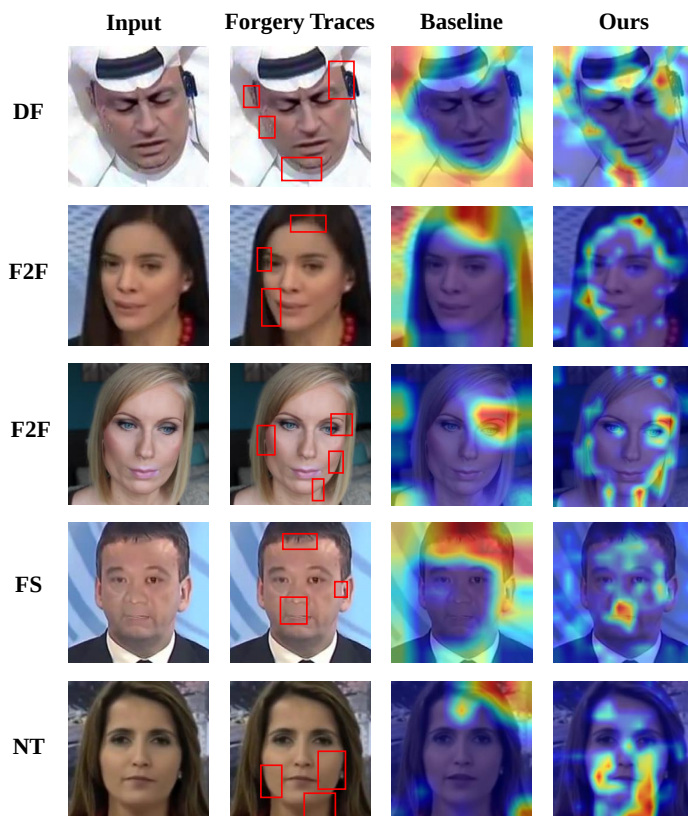


Figure 6: 基线 and 我们的 FGL 的显著性图可视化。红色框表示一些明显或细微的伪造痕迹。

模型在不同伪造技术中的泛化能力。这些视觉观察与我们的定量结果一致，为我们的方法的优越性提供了额外的证据。

2.4.3 显著图可视化

我们使用 Grad-CAM [?] 来突出显示我们提出的方法在伪造人脸上关注的区域。图 6 展示了在 FF++ 数据集上的可视化结果，证明了我们方法更关注于细微的伪造线索。需要注意的是，不同的伪造方法会留下不同的痕迹：DF 和 F2F 通常在面部轮廓和边缘上显示出线索，而 FS 和 NT 则强调与面部表情肌肉和动作相关的伪影。我们的可视化结果显示，FGL 可以准确捕捉这些细微的伪造特征，而基线方法通常只识别出明显的操作痕迹。特别是在处理 NT 时，我们的方法可以识别传统方法通常难以检测的细微质感层次差异。这种精细的注意力分布使我们的方法在处理不同的伪造技术时表现出更强的适应性和鲁棒性。

在这项研究中，我们专注于增强检测模型对未知伪造技术的泛化能力。为了应对深伪技术快速演变带来的挑战，我们提出了伪造引导学习（FGL）策略。FGL 使检测模型能够动态适应未知的伪造技术，同时保持稳定的性能，即便伪造技术不断发展。此外，我们整合了频域感知机制（FPM）和自适应伪造关系感知（AFRP）模块，构建了一个双重感知网络（DPNet）。该网络提高了模型捕捉伪造痕迹的能力，并优化了特征交互。通过将 FGL 策略与 DPNet 结合，我们的方法提高了伪造识别的精度。大量的实验结果表明，FGL 在处理未知伪造技术时显著优于现有方法。这凸显了其在真实场景中的深伪检测有效性。在未来的工作中，我们旨在进一步扩展这一框架，以适应更多样和复杂的伪造场景。