arXiv:2601.00789v1 [cs.CV] 2 Jan 2026

Fusion-SSAT：通过特征融合释放自监督辅助任务在泛化深度伪造检测中的潜力

Shukesh Reddy、Srijan Das、Abhijit Das，*IEEE* 高级会员

摘要—本文尝试释放自监督学习作为辅助任务的潜力，以优化通用深度伪造检测这一主任务。为此，我们考察了这些任务训练方案的不同组合，以找出最有效的方案。研究结果表明，融合自监督辅助任务的特征表示，能够为当前问题提供强大的特征表示。这种表示可以充分发挥其潜力，并融合自监督任务与主任务各自独特的表示，从而提升主任务的性能。我们在大量数据集上进行了实验，包括 DF40、FaceForensics++、Celeb-DF、DFD、FaceShifter 和 UADFV。结果显示，与当前最先进的检测器相比，我们的方法在跨数据集评估中具有更强的泛化能力。

索引 Te rms—广义深度伪造检测、自监督学习、辅助任务、局部方向模式 .

1 引言

RAPID 大型语言模型（LLMs）的使用日益增加[47], [19], 深度生成模型[81] 和扩散模型[56] 彻底改变了文本、图像、视频和音频域的内容创作。这些进步使得创建极其逼真的合成媒体变得更加容易，而这些媒体通常难以与真实内容区分。随着深度伪造生成技术的快速发展[30], [70], [55], [9], [49], [13], [48], [4], [62], [54], [2], [26], [27]在人脸交换[62],[73], 面部表情[2], [9], [13],人脸重演[4], [28],[31], [39],说话人脸生成[27], [80], [59], [67],面部属性编辑[15], [42], [77], [71] 等技术能够将源图像或视频中的内容迁移至目标视频。这类伪造对主体身份识别构成了风险，而身份识别已广泛应用于数字支付、视频监控和社交媒体。为缓解这些问题，针对人脸伪造识别的深度伪造检测器研究正在迅速发展。因此，许多研究工作致力于开发利用卷积神经网络和循环神经网络模型检测篡改媒体的方法：XceptionNet[14],EfficientNetB7[63]；基于变换器的模型：视觉变换器（ViT）[20], Swin Transformer[41]。

尽管当前最先进的深度伪造检测器，如 UCF[75], RECCE[6], CORE[46],FFD[16], Face X-Raycite[34], DSP-FWA[35], Capsule[45], 和 XceptionNet[14] 在域内评估中表现出色，但在跨域或跨数据集分析中，其性能却显著下降。

最先进模型[10], [43], [6], [45], [51], [35], [16], [34], [65], [7], [68], [9]往往难以跨数据集泛化，因为它们倾向于学习表层的、数据集特定的模式，而不是学习被操纵内容背后的生成机制。对数据集特定伪影的过度依赖，限制了它们在跨数据集评估中可靠表现的能力。由此导致的性能下降，揭示了模型在学习稳健且可迁移的特征方面存在关键缺陷，而这些特征应能适用于不同域。

尽管这些模型在受控环境或特定域中可能达到很高的准确率 [75], [46],，但它们往往会过拟合训练数据中的特殊线索，例如压缩伪影或内容规律，从而降低其实际应用价值。在现实场景中，深度伪造在生成方法、压缩级别和内容多样性方面差异很大，这些局限性尤为明显，也凸显了对更稳健、泛化能力更强的检测方法的需求。

为缓解上述局限性，我们提出了一种新方法 Fusion-SSAT，该方法由辅助任务重建和作为主任务的深度伪造检测组成。重建任务旨在从给定的掩蔽 RGB 人脸图像中学习局部纹理特征。主任务是用于深度伪造检测二分类的全局特征编码器。我们将辅助任务编码器提取的特征与主检测任务编码器提取的特征进行融合，以结合全局特征和局部特征。这种局部与全局特征的融合显著提升了模型在不同数据集上的深度伪造检测性能，因为从对应 RGB 图像中重建局部纹理模式，会促使编码器识别操纵过程中产生的复杂人脸伪影与失真，从而增强检测特征表示。具体而言，这一显著更为艰难的特征学习过程促使模型获取对操纵具有鲁棒性的表征。因此，它起到了隐式正则化的作用，推动模型更有效地检测局部线索以及全局线索，而真实与伪造内容之间通常存在差异的正是这些线索。

总而言之，我们的贡献如下： :

• *S. Reddy*、*A. Das* 任职于印度皮拉尼比尔拉理工学院计算机科学与信息系统系的机器智能组。*S. Das* 任职于美利坚合众国北卡罗来纳大学夏洛特分校。电子邮件：*abhijit.das@hyderabad.bits-pilani.ac.in*

- 我们提出了一种新颖的方法 Fusion-SSAT，以提升跨域深度伪造检测场景中的深度伪造检测性能和泛化能力。
- 展示 Fusion-SSAT 如何通过对应人脸图像中的局部纹理模式，学习细粒度纹理和伪影模式。
- 在 DF40 和深度伪造基准数据集上进行了广泛评估，验证了所提出方法的有效性。

2 相关工作

在深度伪造检测中，卷积神经网络通常比其他架构模型更受青睐。早期的深度伪造检测研究主要采用已知的卷积神经网络架构，例如 exception-net[14] 和 EfficientNet B4[63]，作为二元分类器。尽管如此，这些技术在泛化到未见过的操纵方法时仍面临困难。为解决这一问题，研究者探索了多种方案，包括解耦学习、多任务学习[18]，[82]，[3]，和伪造样本合成[44]，[61]。生成建模的进步使深度伪造变得越来越逼真，导致局部误差难以察觉。前述方法容易过时，因为它们仍依赖传统卷积神经网络[14]，[63]，[58]，，因此随着卷积层不断加深，局部信息会逐渐丢失。为增强对低层特征的捕捉能力，有研究提出采用隐式注意力策略的技术[65] ，但这些技术的泛化能力仍然不足。

视觉变换器（ViTs）[20]，[17]，[32]，[3] 的最新进展促成了多种适用于不同应用的类型。要使这些模型达到最佳性能，需要全面的数据集和预训练。DeiT 模型[65]通过先进的正则化、数据增强以及从卷积层提取令牌，得以改进，从而降低了这些要求。T2T[79]采用的令牌化方法旨在识别并记录局部结构信息。其他模型则采用卷积滤波器来融入归纳偏置。分层变换器[8]通过图块合并减少令牌数量，从而融入归纳偏置。尽管如此，它们在预训练时仍需要规模相当大的数据集。

自监督学习已有效应用于基于图像的方法，例如 SimCLR[10]，、MoCo[24]，、MAE[23]，和 DINO[7]，，也应用于基于视频的方法，例如 CoCLR[22]，、VideoMAE[64]，和 SVT[52]。自监督学习注重利用数据的内在结构获取表征，因此无需依赖标注数据。这种方法显著提升了泛化能力；目前的研究在零样本测试中的表现已超过监督模型。借助这些进展，我们认为自监督学习将帮助模型获得更优的表征，从而提升深度伪造检测的泛化能力。基于自监督学习的方法能够使模型准确区分真实样本与伪造样本，不受所采用操纵策略的影响。

多任务学习已在机器学习和深度学习应用的多个领域中得到广泛研究，并已应用于自然语言处理领域，包括统一表征和表征学习。此外，多任务学习还应用于语音识别、药物发现以及计算机视觉领域的应用，例如人脸分析、行人识别、人脸对齐和属性预测等。此外，多任务学习近来在人脸属性学习中日益受到重视，这类学习有时也被称为语义特征，因为它们能够更真实地描述对象和动作。这种方法通过同时对人脸相关活动的多个方面进行建模，有助于全面理解视觉领域。尽管多任务学习通常在监督学习框架下进行研究，但在自监督学习框架下对其的研究仍基本处于空白状态，因此与当前问题尤为相关。此外，值得指出的是，深度伪造检测尚未在多任务学习框架下得到充分研究。

3 提出的方法 3.1 预备知识

自监督辅助任务（SSAT）[18]：在训练数据有限时，联合优化视觉变换器，使其同时执行主任务（分类）和自监督辅助任务（重建）。主任务利用完整图像的潜在表征进行分类，而掩码图像则用于重建任务。该框架通过主任务损失函数 L_{cls} 和辅助任务损失函数 L_{SSAT}进行联合训练，其中 λ 设为 0.1，从而在训练期间为 L_{SSAT} 赋予 0.9 的较高相对权重，如公式 1所示。图1概述了该框架。

$$L = \lambda \cdot L_{\text{cls}} + (1 - \lambda) \cdot L_{\text{SSAT}} \tag{1}$$

SSAT 在高质量深度伪造样本上表现出更强的泛化能力，但无法从高度压缩的深度伪造图像中学习局部和全局特征。

局部纹理特征：编码面部区域中细粒度的空间变化和方向变化，因此能够抵御光照变化、噪声和图像压缩。通过分析较大邻域内的边缘响应，这些描述符能够保留眼睛、鼻子和嘴巴等关键区域周围复杂的局部结构。先进的采样和方向加权增强了其表征细微结构差异的能力，而这些差异可能会被全局描述符忽略。因此，局部纹理特征可作为一种强大的工具，应用于需要精确区分细微且细粒度视觉模式的场景。

3.2 Fusion-SSAT

与最先进的深度伪造检测器不同，仅通过分类任务进行训练会限制模型理解深度伪造伪影各种模式的能力。在 Fusion-SSAT 中，我们引入了一种多任务学习框架：RGB 视频作为深度伪造检测主任务（即真实/伪造分类）的输入，以监督方式执行；而 LDP 视频/局部纹理模式则作为辅助任务进行重建，以自监督方式执行，同时两个任务共享特征。为增强对深度伪造的泛化能力，我们使用来自辅助任务编码器的特征与 RGB 的全局特征组合来训练分类器，与仅使用全局特征进行训练相比，这有助于提升模式识别能力。

令 $V = \{v_1, v_2, \ldots, v_n\}$ 表示一个视频序列数据集，其中每个 v_i 都是由一系列帧组成的视频，并组织为张量 $v_i \in \mathbb{R}^{B \times T \times C \times H \times W}$。这里，$B$表示批量大小，$T$ 表示帧数，C 表示

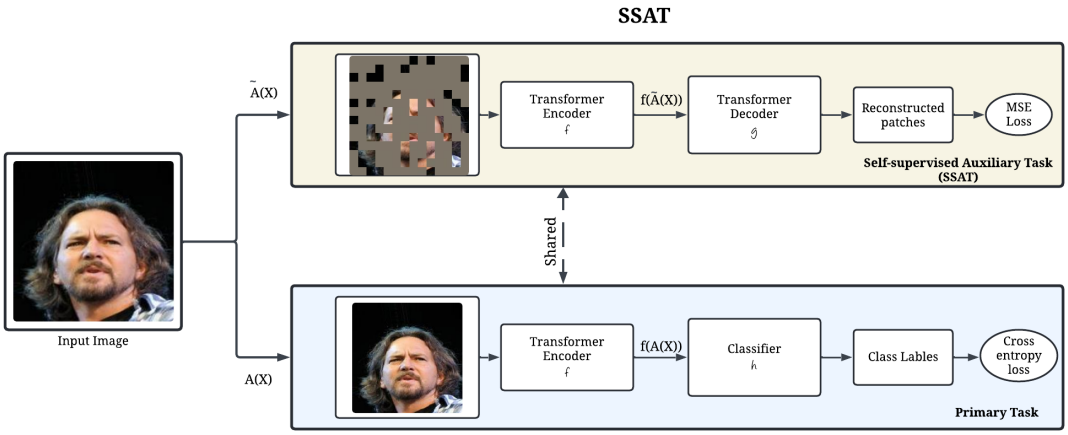


图 1：自监督辅助任务（SSAT）框架概览。

通道数， H 表示帧高， W 表示帧宽。每个视频均以两种形式提供：掩码后的 RGB 视频$L(v_i)$ 和完整图块 RGB 视频 $R(v_i)$，二者的维度均为$\mathbb{R}^{B\times T\times C\times H\times W}$。RGB 视频经过随机掩码~，掩码比例为 0.75，记作 $L(v_i) = M(L(v_i))$，其中 M 会随机掩盖 75% 的时空图块。

Both$(\tilde{v_i})$ 和 $R(v_i)$ 由共享的 ViT 编码器处理f：$\mathbb{R}^{B\times T\times C\times H\times W} \to \mathbb{R}^{B\times S\times D}$，其中 S 表示编码后的词元数量， D 表示潜在表征的维度。编码器生成潜在表征：

$$f(R(v_i)) \in \mathbb{R}^{B\times S\times D} \quad \text{for the RGB video,}$$
$$f'(\tilde{L}(v_i)) \in \mathbb{R}^{B\times S'\times D} \quad \text{for the masked RGB video,}$$

其中，由于掩码操作， $S' \le S$； f' 表示对掩码输入进行操作的编码器。

深度伪造分类的主任务 RGB 视频的潜在表征 $f(R(v_i))$ 被送入分类器 h： $\mathbb{R}^{B\times S\times D} \to \mathbb{R}^{B\times 2}$，用于进行二元深度伪造分类。每个视频 v_i 都对应一个真实标签 $y_i \in \{0,1\}$，其中 $y_i = 1$ 表示深度伪造视频， $y_i = 0$ 表示真实视频。分类损失通过交叉熵计算：

$$L_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B}[y_i \log(h(f(R(v_i)))_1) + (1-y_i)\log(h(f(R(v_i)))_0)]$$

其中， $h(f(R(v_i)))_k$ 表示类别 $k \in \{0,1\}$ 的预测概率。

辅助任务：掩码 LDP2Video 预测任务 掩码 RGB 视频的潜在~表征 $f'(L(v_i))$ 被传入浅层解码器 g：$\mathbb{R}^{B\times S'\times D} \to \mathbb{R}^{B\times T\times C\times H\times W}$，该方法借鉴了 VideoMAE 方法[64]。解码器从原始 RGB $L(v_i)$ 掩码视频~表征中重建局部纹理图块。遵循 MAE 框架， g 以 $f'(L(v_i))$和可学习掩码标记作为输入。解码器输出中的每个标记都经过线性投影，映射为一个表示图像块像素值的向量。重建后的视频为：

$$\hat{L}(v_i) = g(f'(\tilde{L}(v_i))).$$

重建损失是仅在掩码图块上计算的均方误差（MSE）：

$$L_{\text{rec}} = \frac{1}{|\mathcal{M}_i|}\sum_{(t,h,w)\in\mathcal{M}_i}\left\|L(v_i)[:,t,:,h,w] - \hat{L}(v_i)[:,t,:,h,w]\right\|_2^2$$

~其中， $\mathcal{M}_i$ 是 $L(v_i)$中掩码图块索引的集合， $|\mathcal{M}_i|$是掩码图块的数量。

该框架通过对损失函数进行凸组合，同时优化主要分类任务和辅助重建任务：

$$L = \lambda * L_{\text{cls}} + (1-\lambda) * L_{\text{rec}}$$

其中， $\lambda = 0.1$，分类损失的权重设为 0.1，重建损失的权重设为 0.9，以突出辅助任务。融合策略：本文引入联合特征嵌入策略，将来自 LDP[29] 和 RGB 的特征分别表示为$f'(L(v_i))$ 和 $f(R(v_j))$，并通过逐元素~相乘将其组合： $z = f'(L(v_i)) \odot f(R(v_j))$。随后将融合特征向量 z 输入分类器 h。这种联合嵌入使分类器能够同时利用 LDP 特征[29]中的细粒度纹理信息和 RGB 输入中的高层语义线索，因此，与早期的独立处理方法相比，分类性能得以提升，如图 2 所示。该方法确保共享的视觉变换器编码器[20] 能够从 RGB 视频中学习稳健的深度伪造检测表征，同时利用自监督学习重建 LDP 视频中的掩码图块，从而增强泛化能力，并捕捉对深度伪造分析至关重要的基于纹理的特征。

4 实验结果

本节介绍所提出的模型 Fusion-SSAT 的实验细节和结果，并将其与最先进的深度伪造检测器进行比较。第 4.1 节介绍用于评估所提出模型的数据集详情；第 4.2 节介绍实现细节、硬件以及用于报告所提出模型与最先进模型性能的评估指标。第 4.3 节比较所提出模型与现有最先进模型的结果。

4.1 数据集

我们在 FaceForensics++[57]、Celeb-DF v1[36]， Celeb-DF v2[36]，FaceShifter[33],深度伪造检测[57], UADFV[78], DF40[74]数据集上评估了所提出的模型 Fusion-SSAT。所有提出的模型均在 FF++ c23 上训练，并进行了测试

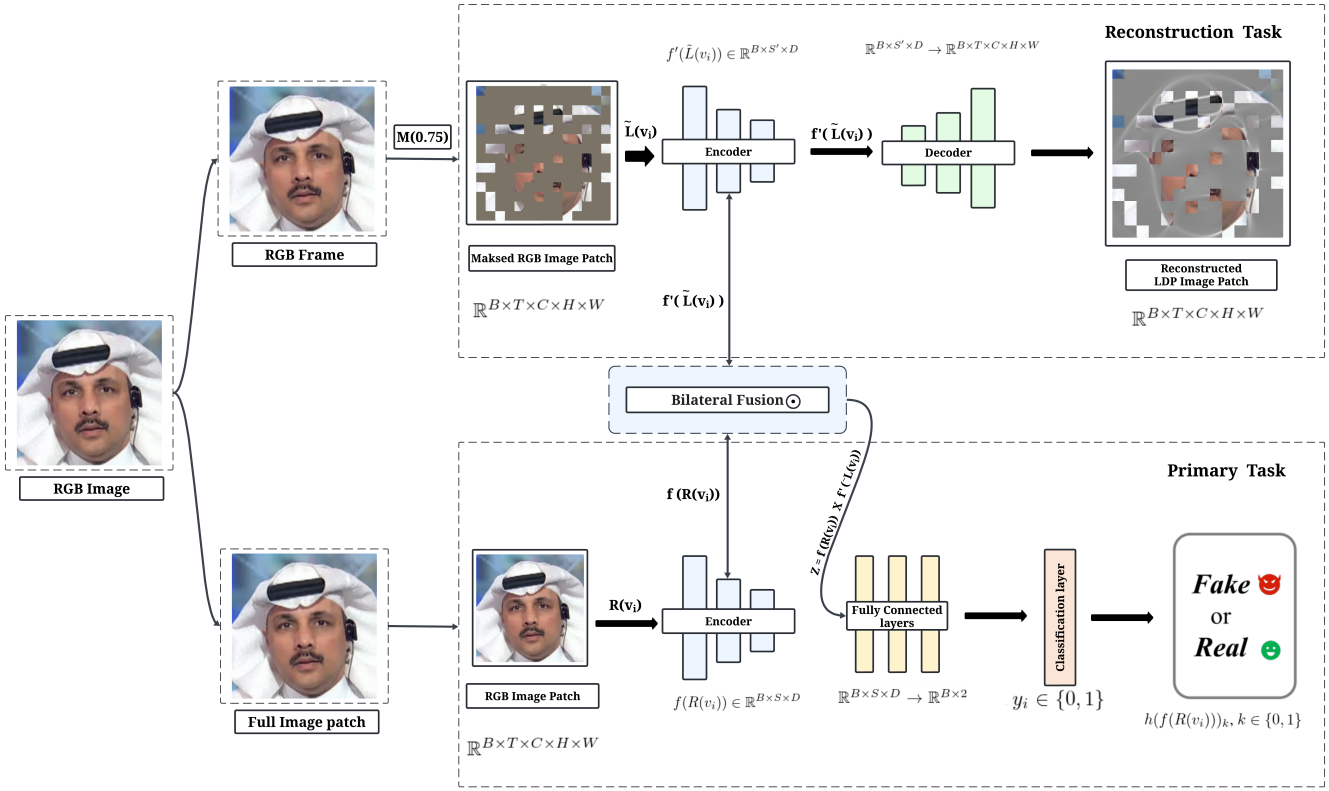


图 2：所提出的 Fusion-SSAT 方法概述

测试所用的数据集包括上述跨域数据集。各数据集的详细介绍如下：

FaceForensics++[57]： 是一个大型基准数据集，包含 1000 个真实视频和 4000 个伪造视频，其中 720 个用于训练集，测试集和验证集各使用 140 个。该数据集包含采用最先进方法生成的篡改视频，即 Face2Face、FaceSwap、DeepFakes 和 Neuraltextures。FF++ 有三个版本，分别对应不同的视频压缩级别：原始、轻度压缩（c23）和高度压缩（c40）。

Celeb-DF v1, v2[36]： 包含高质量换脸视频；Celeb-DF v1 有 408 个原始视频和 795 个伪造视频，而 v2 有 590 个原始视频和 5639 个伪造视频。

Faceshifter[33]： 是一种用于生成高保真且具备遮挡感知能力的换脸视频的方法。它是 FaceForensics++数据集的一个子集，包含 1000 个伪造视频。

深度伪造检测[57]： 是一个由 Google 开发的数据集，包含 363 个真实视频和 3000 个伪造视频。目前，该数据集已纳入 FaceForensics++ 数据集。**UADFV[78]：** 数据集由 98 个视频组成，其中 49 个是真实视频，49 个是伪造视频。

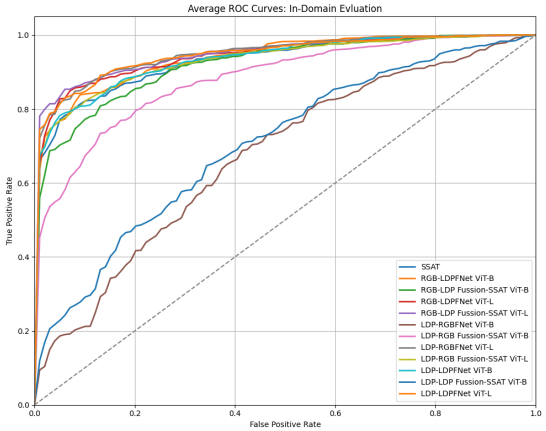
DF40[74]： 是一个大型数据集，汇集了多种先进的深度伪造生成技术，包含 0.1M+ 个伪造视频。该数据集使用 FaceForensics++和 Celeb-DF 数据集中的真实视频来生成伪造视频。数据集大致分为四类：换脸（FS）、人脸重演（FR）、整脸合成（EFS）和人脸编辑（FE）。DF40 由 40 种深度伪造技术构成，包括 10 种换脸方法、13 种人脸重演方法、12 种整脸合成方法和 5 种人脸编辑方法。

4.2 实现细节与评估指标

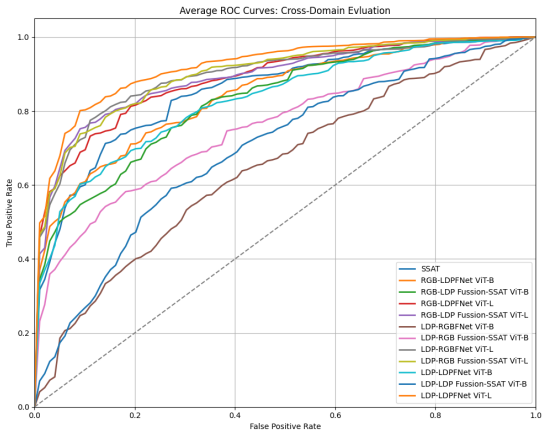
所有提出的模型均使用尺寸为 224 x 224 像素的图像/视频进行训练，并将 16 x 16 像素的图像块作为相应模型任务（分类或重建）的输入（LDP 或 RGB）。训练时，批量大小为 8，在 NVIDIA A100 GPU 上使用随机梯度下降优化器调整权重。学习率采用余弦调度，从 0.00005 开始，最低降至 1e-6。在模型经历的 75 个训练轮次中，学习率始终以 0.05 的衰减率逐步降低。随机深度的初始路径丢弃率设为 0.01，并使用 Lambda（λ）值 0.1 来平衡不同的损失项。掩码比例设为 0.75。所有数据集均使用官方划分。

4.3 结果与讨论

我们在两种不同的场景下评估了所提出的模型 Fusion-SSAT。在第 4.3.1 节中，所有模型均在 FaceForensics++ c23[57]，上训练，并在域内数据集和跨域数据集上进行测试，其中包括 Celeb-DF v1[36]，、Celeb-DF v2[36]，、DFD[57]，、FaceShifter[33]，和 UADFV[78]。在第 4.3.2 节中，所有模型均使用 FF++ 域的 DF40 训练集进行训练，该训练集包含 FS（FF）、FR(FF)、EFS(FF)伪造数据，并在所有伪造方法的伪造方法内测试、跨伪造（FF）[57]和跨域（CDF）[36]场景下进行评估。在我们提出的模型 Fusion-SSAT 的命名规则中，连字符前的模态表示提供给重建任务的掩码视频输入，而连字符后的模态表示进行预测的域。例如，在 RGB-LDP Fusion-SSAT 中，掩码视频输入来自 RGB 模态，而



(a) 在FF++ 数据集上的域内评估



(b) 在CDF、DFD、UADFV上的跨域评估数据集

图3：域内和跨域评估的ROC曲线。

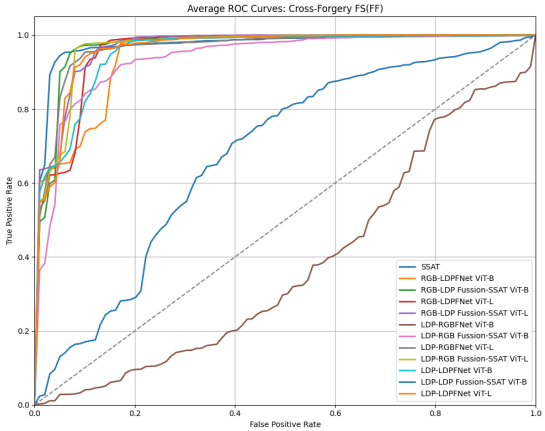
模型则在 LDP 中进行预测。我们采用曲线下面积作为评估指标。我们模型取得的最佳曲线下面积分数用下划线标出，最优的最先进水平结果用斜体表示，而每列性能最高的方法用**粗体**表示。

4.3.1 域内与跨域实验结果

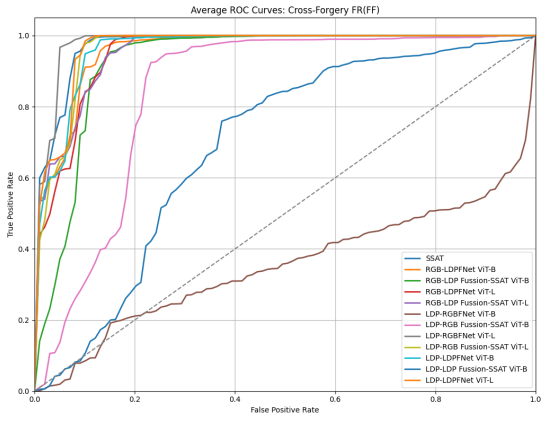
表 1 展示了我们提出的模型在 FF++ c23[57] （深度伪造（FF-DF）、Face2Face（FF- F2F）、FaceSwap（FF-FS）、Neuraltextures（FF-NT））上训练，并在域内评估和跨域数据集上进行评估的实验结果。我们将实验结果与 DeepfakeBench 的[76] 空间检测器的结果进行了基准比较，其性能优于现有最先进的深度伪造检测器。

第 1 列定义了检测器的规格，第 2 列定义了相应检测器所采用的骨干网络。第 3、4 列定义了域内评估和跨域评估的实验结果（曲线下面积得分），其中包括完整数据集 FF-c23[57],FF-c40[57], CDFv1[36],CDFv2[36], DFD[57],Fsh [33],UADFV[78]以及特定伪造方法上的测试得分。“Avg”列概述了每个检测器在所有数据集上的性能，包括域内和跨域性能。

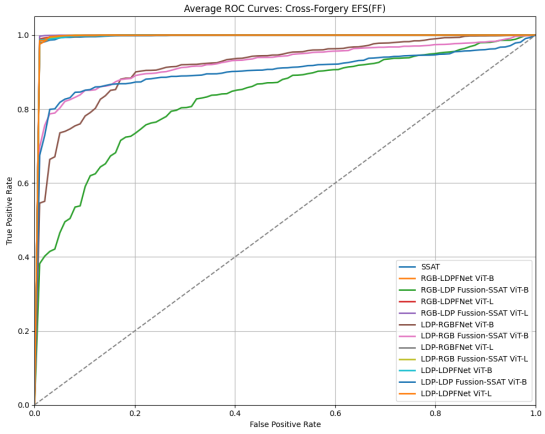
我们提出的模型 Fusion-SSAT，与表1中列出的所有



(a) FS（FF）



(b) FR (FF)



(c) EFS (FF)

图4：DF40数据集上跨伪造评估的ROC曲线。

在上述最先进检测器中，UCF 的平均值最高，达到 0.9527%；然而，我们的方法 Fusion-SSAT 达到了 0.9613% 的 AUC得分，表明在域内评估中比现有最先进方法高出 2%。我们的模型 Fusion-SSAT 在对每种伪造方法进行单独评估时均表现出优越性能，具体包括 FF-c23、FF-DF、FF-F2F、FF-FS 和 FF-NT。在识别压缩深度伪造 FF-c40 时，我们的模型 Fusion-SSAT 的性能下降了 0.005%。在跨域评估中，每个数据集的性能均提升了 5%，平均而言，我们的 AUC 提高了 8%

表 1： 域内和跨域评估，采用曲线下面积指标。所有检测器均在 *FF-c23* 上进行训练，并在其他数据域内和跨域评估，以及总体结果。
“平均值”表示以下各项的平均曲线下面积

检测器	骨干网络	域内评估							跨域评估					
		FF-c23	FF-c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg.	CDFv1	CDFv2	DFD	Fsh	UADFV	Avg.
Meso4 [1]'WIFS-18	MesoNet[1]	0.6077	0.5920	0.6771	0.6170	0.5946	0.5701	0.6097	0.7358	0.6091	0.5481	0.5660	0.7150	0.6348
MesoIncep [1]'WIFS-18	MesoNet[1]	0.7583	0.7278	0.8542	0.8087	0.7421	0.6517	0.7571	0.7366	0.6966	0.6069	0.6438	0.9049	0.7177
CNN-Aug [69]'CVPR-20	ResNet[25]	0.8493	0.7846	0.9048	0.8788	0.9026	0.7313	0.8419	0.7420	0.7027	0.6464	0.5985	0.8739	0.7127
Xception [57]'ICML-19	Xception[14]	0.9637	0.8261	0.9799	0.9785	0.9833	0.9385	0.9450	0.7794	0.7365	0.8163	0.6249	0.9379	0.7790
EfficientB4 [63]'ICCV-19	Efficient[63]	0.9567	0.8150	0.9757	0.9758	0.9797	0.9308	0.9389	0.7909	0.7487	0.8148	0.6162	0.9472	0.7835
Capsule[45]'CASSP-19	Capsule[45]	0.8421	0.7040	0.8669	0.8634	0.8734	0.7804	0.8217	0.7909	0.7472	0.6841	0.6465	0.9078	0.7553
FWA[35]'CVPRW-19	Xception[14]	0.8765	0.7357	0.9210	0.9000	0.8843	0.8120	0.8549	0.7897	0.6680	0.7403	0.5551	0.8539	0.7214
X射线[34]'CVPR-20	HRNet[68]	0.9592	0.7925	0.9794	0.9872	0.9871	0.9290	0.9391	0.7093	0.6786	0.7655	0.6553	0.8989	0.7415
FFD[16]'CVPR-20	Xception[14]	0.9624	0.8237	0.9803	0.9784	0.9853	0.9306	0.9434	0.7840	0.7435	0.8024	0.6056	0.9450	0.7761
CORE[46]'CVPRW 2020	Xception[14]	0.9638	0.8194	0.9787	0.9803	0.9823	0.9339	0.9431	0.7798	0.7428	0.8018	0.6032	0.9412	0.7738
RECCE[6]'CVPR-20	Designed	0.9621	0.8190	0.9797	0.9779	0.9785	0.9357	0.9422	0.7677	0.7319	0.8119	0.6095	0.9446	0.7731
F3Net [50]'ECCV 2020	Xception[14]	0.9635	0.8271	0.9793	0.9796	0.9844	0.9354	0.9449	0.7769	0.7352	0.7975	0.5914	0.9347	0.7614
SPSL [40]'CVPR-21	Xception[14]	0.9610	0.8174	0.9781	0.9754	0.9829	0.9299	0.9408	0.8150	0.7650	0.8122	0.6437	0.9424	***0.7956***
SRM [43]'CVPR-21	Xception[14]	0.9576	0.8114	0.9733	0.9696	0.9740	0.9295	0.9359	0.7926	0.7552	0.8120	0.6014	0.9427	0.6222
UCF[75]'ICCV 2023	Xception[14]	0.9705	0.8399	0.9883	0.9840	0.9896	0.9441	***0.9527***	0.7793	0.7527	0.8074	0.6462	0.9528	***0.7877***
MARLIN [5]'CVPR-23	ViT-L	0.9370	0.8059	0.9399	0.9250	0.9580	0.9708	0.9228	0.7140	0.7030	0.6850	0.6671	0.8150	0.7168
AVAD [21]'CVPR-23	Designed	0.9410	0.7980	0.9872	0.9534	0.9267	0.9093	0.9193	0.7382	0.7089	0.6892	0.7120	0.8910	0.7479
DF-Adapter [60]'IJCV-24	ViT/ViT-Adapter[20]	0.9862	0.9683	0.9984	0.9976	0.9933	0.9597	***0.9839***	0.7174	0.7071	0.7194	0.7292	0.8948	0.7536
StyleGRU [12]'CVPR-24	StyleGRU[12]	0.8297	0.7091	0.9400	0.6850	0.8880	0.8060	0.8297	0.6250	0.6051	0.7220	0.9070	0.8145	0.7347
公平性 [38]'CVPR-24	Xception[14]	0.9828	0.8091	0.9905	0.9865	0.9923	0.9635	0.9541	0.7442	0.7002	0.8482	0.8342	0.8523	0.7958
OPR [11]'NeurIPS-24	EfficientNetB4[63]	0.9591	0.6078	0.9861	0.9272	0.9272	0.9071	0.8858	0.9094	0.8448	0.8091	0.7190	0.8581	0.8281
PFF [37]'ICLR 2024	Xception[14]	0.8518	0.6982	0.8591	0.9281	0.8271	0.8561	0.8367	0.7721	0.7041	0.7826	0.6982	0.8125	0.7539
SSAT [18]'WACV-24	VideoMAE/ViT-B	0.9682	0.7438	0.9991	0.9603	0.9922	0.9215	0.9308	0.8073	0.7757	0.8109	0.9902	0.8920	0.8552
LDP-RGBFNet[53]'ICPRW-24	VideoMAE/ViT-B	0.6766	0.6566	0.8622	0.6047	0.6112	0.6281	0.6732	0.4607	0.6828	0.5654	0.8052	0.7344	0.6497
LDP-RGBFNet[53]的 ICPRW-24	VideoMAE/ViT-L	0.9868	0.7687	0.9998	0.9871	0.9972	0.9631	0.9505	0.8256	0.8165	0.8973	0.9933	0.9545	0.8974
LDP-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-B	0.9687	0.7727	0.9995	0.9667	0.9893	0.9193	0.9360	0.7676	0.7499	0.7537	0.9915	0.8971	0.8319
LDP-LDPFNet[53]'ICPRW-24	VideoMAE/ViT-L	0.9857	0.7673	0.9996	0.9844	0.9969	0.9617	0.9493	0.9193	0.8676	0.8766	0.9946	0.9517	0.9219
RGB-LDPFNet[53]'ICPRW-24	VideoMAE/视觉变换器-B	0.9740	0.7724	0.9996	0.9688	0.9914	0.9364	0.9404	0.8055	0.7550	0.7634	0.9904	0.9077	0.8444
RGB-LDPFNet[53]的ICPRW-24	VideoMAE/视觉变换器-L	0.9815	0.8283	0.9996	0.9828	0.9940	0.9496	0.9559	0.8727	0.8477	0.7955	0.9902	0.9409	0.8894
RGB-LDP Fusion-SSAT	VideoMAE/视觉变换器-B	0.9491	0.7796	0.9941	0.9451	0.9866	0.8710	0.9209	0.7170	0.7659	0.7551	0.9807	0.9311	0.8299
RGB-LDP Fusion-SSAT	VideoMAE/ViT-L	0.9865	**0.8349**	0.9996	**0.9921**	**0.9978**	0.9571	0.9613	0.8740	0.8256	0.8122	**0.9962**	**0.9570**	0.8930

相比现有最先进方法的得分，表明我们的模型 Fusion-SSAT 对未见过的深度伪造操纵技术表现出更优的泛化能力。ROC曲线见图 3。

4.3.2 跨伪造和域上的实验结果表2 总结了我们提出的模型在 DF40 FF域上进行训练，并在伪造内、跨伪造和跨域数据上进行评估的实验结果，其中包括 FS(FF)、FR(FF) 和 EFS(FF)。第1列指定用于模型训练的伪造技术，第2列定义用于评估的模型，第3列展示 DF40测试集（FF域和 CDF域）上的实验结果。

跨伪造评估：我们提出的模型Fusion-SSAT优于所有现有模型，在FS（FF）上的平均曲线下面积为0.970%，在FR (FF)上为0.958%，在EFS（FF）上为0.999%。在直接比较中，我们的模型在曲线下面积方面呈现出显著的上升趋势：在FF（FF）上训练时提升了10%，在FR (FF)上训练时提升了12%，在EFS（FF）上训练时最高提升了30%。

尽管所有最先进模型在EFS测试数据上的表现都较差，但我们的模型在FS和FR测试集上展现出更好的泛化能力。在EFS（FF）训练集上训练时，最先进模型在所有测试场景中的性能下降幅度最高达50%。相比之下，我们提出的方法始终优于这些模型，在FS（FF）和FR (FF)测试集上均实现了0.99%的曲线下面积。这些实验的ROC曲线见图4。

跨域评估： 如表 2所示，当在 FF域上进行训练和测试时，大多数当前最先进水平的深度伪造检测模型表现优异，但在CDF）跨域数据上进行测试时，性能较差，如表2所示。

我们提出的模型 Fusion-SSAT 在 FS(CDF)、FR(CDF) 和 EFS(CDF) CDF 测试集上的平均曲线下面积分别达到 0.986%、0.981% 和 0.991%，超越了所列出的所有现有最先进模型。即使在一对一跨域测试中，所有最先进模型的性能最高也仅达到 70% 的曲线下面积，而我们提出的模型的曲线下面积均不低于 89%，这表明与现有深度伪造检测器相比，我们的模型具有显著更强的泛化能力。在 FR (CDF)测试中，大多数现有方法表现出

表2：跨伪造与跨域评估：在 *DF40 FF*域上使用不同伪造方法（*FS*、*FR*、*EFS*）训练的模型在 *FF* 域和 *CDF* 域上进行评估。以及 *CDF*域。

训练集	模型	测试集（FF）				测试集（CDF）			
		FS	FR	EFS	平均（FF）	FS	FR	EFS	平均（CDF）
FS（FF）	Xception[14]'CVPR-17	0.991	0.892	0.810	0.898	0.922	0.657	0.642	***0.740***
	CLIP[51]'ICML-21	0.996	0.908	0.837	***0.914***	0.967	0.744	0.730	0.814
	SRM[43]'CVPR-21	0.988	0.867	0.703	0.853	0.919	0.621	0.603	0.714
	SPSL[40]'CVPR-21	0.987	0.849	0.735	0.857	0.938	0.656	0.648	***0.747***
	RECCE[6]'CVPR-20	0.991	0.855	0.758	0.868	0.926	0.632	0.610	0.723
	RFM[66]'CVPR-21	0.992	0.884	0.821	***0.899***	0.939	0.637	0.628	0.735
	SSAT [18]'WACV-24	0.988	0.963	0.972	0.974	0.985	0.860	0.990	0.945
	LDP-RGBFNet 视觉变换器-B[53]'ICPRW-24	0.458	0.354	0.322	0.378	0.577	0.432	0.711	0.573
	LDP-RGBFNet 视觉变换器-L[53]'ICPRW-24	0.991	0.950	0.954	0.965	0.994	0.882	0.992	0.956
	LDP-LDPFNet 视觉变换器-B[53]'ICPRW-24	0.991	0.978	0.899	0.956	0.990	0.895	0.968	0.951
	LDP-LDPFNet 视觉变换器-L[53]'ICPRW-24	0.993	0.970	0.932	0.965	0.995	0.916	0.988	0.966
	RGB-LDPFNet 视觉变换器-B[53]'ICPRW-24	0.994	0.980	0.870	0.948	0.992	0.913	0.976	0.960
	RGB-LDPFNet ViT-L[53]'ICPRW-24	0.994	0.981	0.905	0.960	0.998	0.951	0.979	0.976
	RGB-LDP Fusion-SSAT ViT-B	0.978	0.982	0.950	0.970	0.995	0.949	0.989	0.977
	RGB-LDP Fusion-SSAT ViT-L	**0.996**	**0.993**	0.920	0.970	**0.998**	**0.976**	0.984	0.986
FR (FF)	Xception[14]'CVPR-17	0.838	0.996	0.670	0.835	0.481	0.857	0.369	0.569
	CLIP[51]'ICML-21	0.932	0.999	0.798	***0.910***	0.638	0.933	0.209	***0.593***
	SRM[43]'CVPR-21	0.893	0.998	0.698	0.863	0.454	0.869	0.326	0.550
	SPSL[40]'CVPR-21	0.901	0.998	0.695	0.865	0.479	0.852	0.256	0.529
	RECCE[6]'CVPR-20	0.865	0.997	0.716	0.859	0.452	0.881	0.332	0.555
	RFM[66]'CVPR-21	0.892	0.999	0.776	***0.889***	0.492	0.882	0.359	***0.578***
	SSAT [18]'WACV-24	0.991	0.997	0.937	0.975	0.968	0.982	0.955	0.968
	LDP-RGBFNet 视觉变换器-B[53]'ICPRW-24	0.379	0.661	0.017	0.352	0.348	0.550	0.096	0.331
	LDP-RGBFNet 视觉变换器-L[53]'ICPRW-24	**0.996**	**0.996**	0.955	0.982	**0.994**	0.994	0.980	0.989
	LDP-LDPFNet ViT-B[53]'ICPRW-24	0.980	0.993	0.919	0.964	0.970	0.982	0.932	0.961
	LDP-LDPFNet 视觉变换器-L[53]'ICPRW-24	0.988	0.993	0.924	0.968	0.991	0.992	0.966	0.983
	RGB-LDPFNet ViT-B[53]'ICPRW-24	0.972	0.987	0.914	0.958	0.972	0.979	0.962	0.971
	RGB-LDPFNet 视觉变换器-L[53]'ICPRW-24	0.975	0.985	0.894	0.951	0.981	0.980	0.949	0.970
	RGB-LDP Fusion-SSAT ViT-B	0.936	0.954	0.894	0.928	0.964	0.973	0.947	0.961
	RGB-LDP Fusion-SSAT ViT-L	0.984	0.994	0.890	0.956	0.990	**0.995**	0.959	0.981
EFS（FF）	Xception[14]'CVPR-17	0.665	0.807	0.999	0.824	0.586	0.594	0.983	0.721
	CLIP[51]'ICML-21	0.688	0.889	0.999	***0.859***	0.617	0.735	0.988	***0.780***
	SRM[43]'CVPR-21	0.596	0.776	0.999	0.790	0.589	0.620	0.964	0.724
	SPSL[40]'CVPR-21	0.659	0.811	0.999	0.823	0.635	0.651	0.975	0.754
	RECCE[6]'CVPR-20	0.691	0.801	0.999	***0.830***	0.623	0.603	0.984	0.737
	RFM[66]'CVPR-21	0.653	0.795	0.999	0.816	0.644	0.666	0.981	***0.764***
	SSAT [18]'WACV-24	0.997	0.997	1.0	0.998	0.995	0.959	1.0	0.985
	LDP-RGBFNet 视觉变换器-B[53]'ICPRW-24	0.924	0.837	1.0	0.920	0.823	0.719	1.0	0.847
	LDP-RGBFNet 视觉变换器-L[53]'ICPRW-24	0.999	0.998	1.0	0.999	0.997	0.981	1.0	0.993
	LDP-LDPFNet 视觉变换器-B[53]'ICPRW-24	0.999	0.997	1.0	0.998	0.997	0.979	1.0	0.992
	LDP-LDPFNet ViT-L[53]'ICPRW-24	0.999	0.998	1.0	0.999	0.997	0.981	**1.0**	0.993
	RGB-LDPFNet ViT-B[53]'ICPRW-24	0.999	0.996	0.999	0.998	0.996	0.965	0.990	0.983
	RGB-LDPFNet ViT-L[53]'ICPRW-24	0.999	0.998	0.999	0.998	0.995	0.980	0.990	0.988
	RGB-LDP Fusion-SSAT ViT-B	0.996	0.964	0.999	0.986	0.971	0.840	0.990	0.933
	RGB-LDP Fusion-SSAT ViT-L	**0.999**	**0.999**	0.999	0.999	**0.997**	**0.986**	0.990	0.991
All	MARALIN [5]'CVPR-23	-	-	-	0.981	-	-	-	0.796
	GrDT [72]'WACVW-25	-	-	-	0.986	-	-	-	0.824
	Ours	-	-	-	0.999	-	-	-	0.991

性能下降，平均曲线下面积低于 0.60%。另一方面，我们的模型取得了显著更高的平均曲线下面积，分别达到 0.97% 和 0.98%。在所有当前方法表现不佳的 CDF 域测试中，尤其是在 FR（FF）场景中，平均曲线下面积经常低于 0.60，而我们的模型展现出出色的可靠性，在所有测试集上的曲线下面积均稳定高于 0.940。这些实验的 ROC曲线见图 5。

Fusion-SSAT 在所有测试场景中始终优于其他模型，取得了最高的平均曲线下面积分数。这表明它在不同类型的伪造内容和区域上具有出色的泛化能力。这些发现证实，将基于纹理的特征与全局特征相结合，能够提升跨域深度伪造检测的效果。

4.4 消融研究

作为消融研究的一部分，我们评估了不同特征融合策略的影响，包括 LBP-LBP、LDP-LDP 和 LDP-RGB。表 3、4的结果表明，仅依赖局部纹理特征（LBP-LBP、LDP-LDP）会导致性能不佳。值得注意的是，将局部描述符与全局表示相结合（LDP-RGB），能够得到更具判别性的特征，并提升整体性能。

5 结论

在这项工作中，我们提出了一种名为 Fusion-SSAT 的新方法，将自监督辅助任务的特征表示与主任务融合。我们将掩码LDP局部描述符用作 SSAT任务的输入，并尝试重建 RGB图像块，这使自监督学习任务更难学习局部模式，而不是

表3：在 *FF++[57]*数据集上训练的 *LDP-LDP* 和 *LDP-RGB* 特征融合的域内评估与跨域评估

检测器	骨干网络	域内评估							跨域评估					
		FF-c23	FF-c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg.	CDFv1	CDFv2	DFD	Fsh	UADFV	Avg.
LDP-RGB 融合 SSAT	VideoMAE/ViT-B	0.8833	0.7694	0.9899	0.8163	0.9467	0.7814	0.8645	0.5963	0.7071	0.6500	0.9274	0.9115	0.7585
LDP-RGB 融合 SSAT	VideoMAE/ViT-L	**0.9866**	0.7731	0.9997	0.9887	0.9973	**0.9605**	0.9510	**0.8748**	**0.8254**	**0.8557**	**0.9947**	**0.9540**	0.9009
LDP-LDP 融合 SSAT	VideoMAE/ViT-B	0.7109	0.6696	0.9242	0.5941	0.7051	0.6209	0.7041	0.5309	0.754	0.6157	0.7482	0.8426	0.6983
LDP-LDP 融合 SSAT	VideoMAE/ViT-L	0.9846	**0.7786**	**0.9999**	**0.9849**	**0.9966**	0.9581	0.9505	0.8382	0.8158	0.8058	0.9937	0.9281	0.8763
LBP-RGB 融合 SSAT	VideoMAE/ViT-B	0.8243	**0.7023**	0.9689	0.8262	**0.9052**	0.7325	0.8266	0.5262	0.7001	0.6234	**0.9532**	0.9013	0.7408
LBP-RGB 融合 SSAT	VideoMAE/ViT-L	**0.9123**	0.6512	**0.9781**	**0.9345**	0.8469	0.7934	0.8527	**0.7876**	**0.8011**	**0.7350**	0.8692	**0.9021**	0.8190
LBP-LBP 融合 SSAT	VideoMAE/ViT-B	0.5311	0.6087	0.7420	0.5803	0.6234	0.5698	0.6092	0.4882	0.5405	0.6120	0.5931	0.6598	0.5787
LBP-LBP 融合 SSAT	VideoMAE/ViT-L	0.8243	0.6625	0.9711	0.8442	0.9107	**0.7981**	0.8351	0.7821	0.7940	0.7011	0.8250	0.8792	0.7963

表 4: 跨伪造 以及在 *DF40* 数据集上训练的 *LDP-LDP* 和 *LDP-RGB* 特征融合的跨域评估 *et*

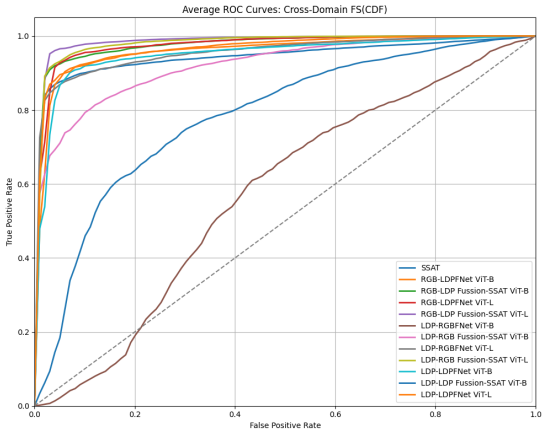
训练集	模型	测试集 (FF)				测试集 (CDF)			
		FS	FR	EFS	平均值（FF）	FS	FR	EFS	平均（CDF）
FS（FF）	LDP-RGB Fusion-SSAT ViT-B	0.948	0.906	**0.955**	0.936	0.940	0.826	**0.995**	0.920
	LDP-RGB Fusion-SSAT ViT-L	0.989	0.980	0.932	0.967	0.995	0.962	0.988	0.982
	LDP-LDP 融合 SSAT ViT-B	0.687	0.620	0.721	0.676	0.774	0.654	0.918	0.782
	LDP-LDP 融合 SSAT ViT-L	**0.991**	**0.985**	0.963	0.979	**0.997**	**0.966**	0.991	0.985
FR (FF)	LDP-RGB Fusion-SSAT ViT-B	0.834	0.895	0.799	0.842	0.865	0.896	0.929	0.897
	LDP-RGB Fusion-SSAT ViT-L	0.983	0.990	0.917	0.963	**0.991**	**0.994**	0.960	0.982
	LDP-LDP 融合 SSAT ViT-B	0.609	0.760	0.70	0.689	0.654	0.726	0.827	0.736
	LDP-LDP 融合 SSAT ViT-L	**0.992**	**0.995**	**0.985**	0.990	0.990	0.992	**0.984**	0.989
EFS（FF）	LDP-RGB Fusion-SSAT ViT-B	0.940	0.836	1.0	0.925	0.767	0.600	1.0	0.789
	LDP-RGB Fusion-SSAT ViT-L	**0.999**	**0.998**	**1.0**	0.999	**0.997**	**0.981**	**1.0**	0.993
	LDP-LDP 融合 SSAT ViT-B	0.929	0.786	1.0	0.905	0.69	0.496	0.999	0.728
	LDP-LDP 融合 SSAT ViT-L	0.999	0.998	**1.0**	0.999	0.997	0.976	0.999	0.991

依赖浅层特征。这提升了主任务的性能，使其能够更好地泛化到未见过的篡改，或用于跨数据集评估。实验结果表明，与最先进的深度伪造检测器相比，该方法在包含最新深度伪造生成技术的大规模 DF40数据集上具有更强的泛化能力。

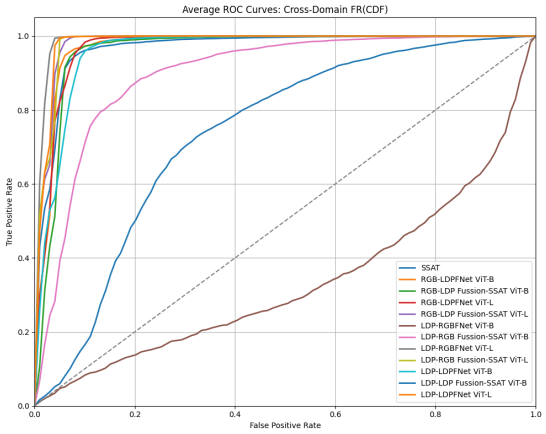
参考文献

[1] D. Afchar、V. Nozick、J. Yamagishi 和 I. Echizen。《MesoNet：一种紧凑的面部视频伪造检测网络》。载于*2018 IEEE*信息取证与安全国际研讨会，第 1–7 页。IEEE，2018。[2] X. An、X. Zhu、Y. Gao、Y. Xiao、Y. Zhao、Z. Feng、L. Wu、B. Qin、M. Zhang、D. Zhang 和 Y. Fu。《部分全连接：在单台机器上训练 1000 万个身份》。载于*ICCVW*，2021。[3] P. Balaji、A. Das、S. Das 和 A. Dantcheva。《通过探索多任务学习，在深度伪造检测过程中关注泛化能力》。载于*2023 IEEE/CVF*计算机视觉国际会议研讨会（*ICCVW*），第 475–484 页，2023。[4] S. Bounareli、C. Tzelepis、V. Argyriou、I. Patras 和 G. Tzimiropoulos。《Hyperreenact：通过联合学习人脸细化与重定向实现单次重演》。载于*IEEE/CVF*计算机视觉国际会议论文集（*ICCV*），2023。[5] Z. Cai、S. Ghosh、K. Stefanov、A. Dhall、J. Cai、H. Rezatofighi、R. Haffari 和 M. Hayat。《MARLIN：用于面部视频表征学习的掩码自编码器》。载于*2023 IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第 1493–1504 页，2023。[6] J. Cao、C. Ma、T. Yao、S. Chen、S. Ding 和 X. Yang。《用于人脸伪造检测的端到端重建-分类学习》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第 4113–4122 页，2022年6月。[7] M. Caron、H. Touvron、I. Misra、H. J´egou、J. Mairal、P. Bojanowski 和 A. Joulin。《自监督视觉变换器的新兴特性》。载于计算机视觉国际会议论文集（*ICCV*），2021。[8] I. Chalkidis、X. Dai、M. Fergadiotis、P. Malakasiotis 和 D. Elliott。《高效长文档分类中层次注意力变换器的探索》，2022。

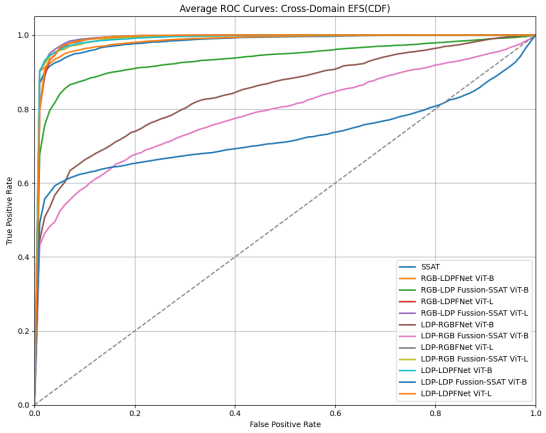
[9] R. Chen、X. Chen、B. Ni 和 Y. Ge。SimSwap：一种高效的高保真人脸交换框架。载于*MM '20*：第*28*届*ACM*多媒体国际会议，2020年。[10] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton。用于视觉表征对比学习的简单框架。*arXiv*预印本 *arXiv:2002.05709*，2020年。[11] J. Cheng、Z. Yan、Y. Zhang、Y. Luo、Z. Wang 和 C. Li。我们能否在训练深度伪造检测器时摆脱深度伪造数据？载于第三十八届神经信息处理系统年会，2024年。[12] J. Choi、T. Kim、Y. Jeong、S. Baek 和 J. Choi。利用风格潜在流实现深度伪造视频检测的泛化。载于*2024*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第1133–1143页，2024年。[13] Y. Choi、M. Choi、M. Kim、J.-W. Ha、S. Kim 和 J. Choo。StarGAN：用于多域图像到图像转换的统一生成对抗网络。载于*IEEE*计算机视觉与模式识别会议论文集，2018年。[14] F. Chollet。Xception：采用深度可分离卷积的深度学习，2017年。[15] Y. Dalva、S. F. Altındis。和 A. Dundar。VecGAN：利用可解释潜在方向实现图像到图像转换。载于欧洲计算机视觉会议，第153–169页。施普林格，2022年。[16] H. Dang、F. Liu、J. Stehouwer、X. Liu 和 A. Jain。关于数字人脸操纵的检测。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第5780–5789页，2020年6月。[17] A. Das、S. Das 和 A. Dantcheva。揭秘用于深度伪造检测的注意力机制。载于*2021*年第*16*届*IEEE*自动人脸与手势识别国际会议（*FG 2021*），第1–7页，2021年。[18] S. Das、T. Jain、D. Reilly、P. Balaji、S. Karmakar、S. Marjit、X. Li、A. Das 和 M. Ryoo。有限数据，无限潜力：关于经掩码自编码器增强的视觉变换器的研究。*2024*年*IEEE/CVF*计算机视觉应用冬季会议（*WACV*），2024年。[19] DeepSeek-AI。DeepSeek-V3技术报告，2025年。[20] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Un- terthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby。一张图像胜过16×16个词：用于大规模图像识别的变换器，2021年。[21] C. Feng、Z. Chen 和 A. Owens。通过视听异常检测进行自监督视频取证。载于2023年IEEE/CVF计算机视觉与



(a) FS（CDF）



(b) FR（CDF）



(c) EFS（CDF）

图 5：DF40数据集上的跨域评估 ROC曲线。

模式识别会议（CVPR），第10491–10503页，2023年。[22] T. Han、W. Xie 和 A. Zisserman。用于视频表征学习的自监督协同训练。载于第34届神经信息处理系统国际会议论文集，NIPS '20，美国纽约州Red Hook，2020年。Curran Associates Inc. [23] K. He、X. Chen、S. Xie、Y. Li、P. Doll´ar 和 R. Girshick。掩码自编码器是可扩展的视觉学习器。arXiv:2111.06377，2021年。[24] K. He、H. Fan、Y. Wu、S. Xie 和 R. Girshick。用于无监督视觉表征学习的动量对比。arXiv预印本 arXiv:1911.05722，2019年。[25] K. He、X. Zhang、S. Ren 和 J. Sun。用于图像识别的深度残差学习。载于IEEE/CVF计算机视觉与模式识别会议论文集，第770–778页，2016年。[26] J. Ho、A. Jain 和 P. Abbeel。去噪扩散概率模型。arXiv预印本 arxiv:2006.11239，2020年。[27] F.-T. Hong、L. Zhang、L. Shen 和 D. Xu。用于说话头视频生成的深度感知生成对抗网络。载于CVPR，2022年。[28] G.-S. Hsu、C.-H. Tsai 和 H.-Y. Wu。双生成器人脸重演。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第642–650页，2022年6月。[29] T. Jabid、M. H. Kabir 和 O. Chae。局部方向模式（LDP）——一种用于目标识别的鲁棒图像描述符。载于2010年第7届IEEE高级视频与信号监控国际会议论文集，AVSS '10，第482–487页，美国，2010年。IEEE计算机学会。[30] A. C. K、A. A V、S. Das 和 A. Das。用于深度伪造视频生成的潜在流扩散。载于IEEE/CVF计算机视觉与模式识别会议研讨会论文集，第3781–3790页，2024年6月。[31] M. Koujan、M. Doukas、A. Roussos 和 S. Zafeiriou。Head2head：基于视频的神经头部合成。载于2020年第15届IEEE自动人脸与手势识别国际会议（FG 2020）（FG），第319–326页，美国加利福尼亚州洛斯阿拉米托斯，2020年5月。IEEE计算机学会。[32] K. Kuckreja、X. Hoque、N. Poddar、S. Reddy、A. Dhall 和 A. Das。Indiface：以全面的合成数据集揭示印度的深度伪造生态。载于2024年第18届IEEE自动人脸与手势识别国际会议（FG），第1–9页，2024年。[33] L. Li、J. Bao、H. Yang、D. Chen 和 F. Wen。推进高保真身份交换以进行伪造检测。载于IEEE/CVF计算机视觉与模式识别会议论文集，第5074–5083页，2020年。[34] L. Li、J. Bao、T. Zhang、H. Yang、D. Chen、F. Wen 和 B. Guo。Face X-Ray：实现更通用的人脸伪造检测。载于2020年IEEE/CVF计算机视觉与模式识别会议（CVPR），第5000–5009页，2020年。[35] Y. Li 和 S. Lyu。通过检测人脸扭曲伪影揭露深度伪造视频。载于IEEE计算机视觉与模式识别会议研讨会（CVPRW），2019年。[36] Y. Li、X. Yang、P. Sun、H. Qi 和 S. Lyu。Celeb-DF：一个大规模且具挑战性的深度伪造取证数据集，2020年。[37] J. Liang、S. Liang、A. Liu、X. Jia、J. Kuang 和 X. Cao。中毒伪造人脸：面向人脸伪造检测的后门攻击。载于第十二届国际学习表征会议，2024年。[38] L. Lin、X. He、Y. Ju、X. Wang、F. Ding 和 S. Hu。在深度伪造检测中保持公平性泛化。载于2024年IEEE/CVF计算机视觉与模式识别会议（CVPR），第16815–16825页，2024年。[39] D. Liu、L. Y. Wu、B. Li、Y. Zhao、Z. Ge 和 J. Zhang。T-person-gan：结合身份一致性与流形混合的文本到人物图像生成。《专家系统与应用》，245：123456，2024年。在线可查。[40] H. Liu、X. Li、W. Zhou、Y. Chen、Y. He、H. Xue、W. Zhang 和 N. Yu。空间相位浅层学习：重新思考频域中的人脸伪造检测。载于2021年IEEE/CVF计算机视觉与模式识别会议（CVPR），第772–781页，2021年。[41] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo。Swin Transformer：利用移位窗口的分层视觉Transformer，2021年。[42] Y. Lu 和 T. Ebrahimi。迈向人工智能合成人脸图像的检测。arXiv预印本 arXiv:2402.08750，2024年。[43] Y. Luo、Y. Zhang、J. Yan 和 W. Liu。利用高频特征实现人脸伪造检测的泛化。载于2021年IEEE/CVF计算机视觉与模式识别会议（CVPR），第16312–16321页，2021年。[44] D. Nguyen、N. Mejri、I. P. Singh、P. Kuleshova、M. Astrid、A. Kacem、E. Ghorbel 和 D. Aouada。Laa-net：用于质量无关且可泛化深度伪造检测的局部伪影注意力网络，2024年。[45] H. H. Nguyen、J. Yamagishi 和 I. Echizen。Capsule-forensics：使用胶囊网络检测伪造图像与视频，2018年。

[46] Y. Ni、D. Meng、C. Yu、C. Quan、D. Ren 和 Y. Zhao。CORE：用于人脸伪造检测的一致表征学习。载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集，第 12–21 页，2022年6月。[47] OpenAI。Gpt-4 技术报告，2024。[48] O. Patashnik、Z. Wu、E. Shechtman、D. Cohen-Or 和 D. Lischinski。Styleclip：文本驱动的 StyleGAN 图像风格操控。载于 IEEE/CVF 国际计算机视觉会议（ICCV）论文集，第 2085–2094 页，2021年10月。[49] K. R. Prajwal、R. Mukhopadhyay、V. P. Namboodiri 和 C. Jawahar。野外语音到唇形生成只需一个唇同步专家。载于第 28 届 ACM多媒体国际会议论文集，MM '20，第 484–492 页，美国纽约州纽约市，2020。计算机协会。[50] 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，第 86–103 页。施普林格，2020。[51] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever。通过自然语言监督学习可迁移的视觉模型，2021。[52] K. Ranasinghe、M. Naseer、S. Khan、F. Khan 和 M. Ryoo。自监督视频变换器。载于 CVPR，第 2864–2874 页，2022年6月。[53] S. Reddy、N. Poddar、S. Das 和 A. Das。用于人脸分析的纹理与模型混合式鲁棒公平特征学习的自监督辅助学习。载于 S. Palaiahnakote、S. Schuckers、J.-M. Ogier、P. Bhattacharya、U. Pal 和 S. Bhattacharya 编，《模式识别：ICPR 2024 国际研讨会与挑战赛》，第 386–401 页，瑞士查姆，2025。施普林格自然瑞士。[54] X. Ren、A. Lattas、B. Gecer、J. Deng、C. Ma 和 X. Yang。通过隐式表征恢复面部几何细节。载于 2023 年 IEEE 第 17 届自动人脸与手势识别国际会议（FG），2023。[55] Y. Ren、G. Li、Y. Chen、T. H. Li 和 S. Liu。Pirenderer：通过语义神经渲染实现可控肖像图像生成，2021。[56] R. Rombach、A. Blattmann、D. Lorenz、P. Esser 和 B. Ommer。利用潜在扩散模型进行高分辨率图像合成，2021。[57] A. R¨ ossler、D. Cozzolino、L. Verdoliva、C. Riess、J. Thies 和 M. Nießner。FaceForensics++：学习检测经过操纵的人脸图像。载于计算机视觉国际会议（ICCV），2019。[58] R. Roy、I. Joshi、A. Das 和 A. Dantcheva。用于深度伪造检测的 3D 卷积神经网络架构与注意力机制，第 213–234 页。施普林格国际出版，瑞士查姆，2022。[59] Z. Sha、Y. Tan、M. Li、M. Backes 和 Y. Zhang。Zerofake：零样本检测由文生图模型生成和编辑的伪造图像。载于 2024 年 ACM SIGSAC 计算机与通信安全会议论文集（CCS '24），美国犹他州盐湖城，2024年10月。ACM。[60] R. Shao、T. Wu、L. Nie 和 Z. Liu。Deepfake-adapter：用于深度伪造检测的双层级适配器。《国际计算机视觉杂志》（IJCV），2024。[61] K. Shiohara 和 T. Yamasaki。利用自混合图像检测深度伪造。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 18720–18729 页，2022。[62] K. Shiohara、X. Yang 和 T. Taketomi。Blendface：重新设计用于换脸的身份编码器。载于 IEEE/CVF 国际计算机视觉会议（ ICCV ）论文集，2023。[63] M. Tan 和 Q. V. Le。Efficientnet：重新思考卷积神经网络的模型缩放，2020。[64] Z. Tong、Y. Song、J. Wang 和 L. Wang。VideoMAE：掩码自编码器是用于自监督视频预训练的数据高效学习器。载于《神经信息处理系统进展》，2022。[65] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablayrolles 和 H. Jegou。训练数据高效的图像变换器及通过注意力进行蒸馏。载于 M. Meila 和 T. Zhang 编，第 38 届机器学习国际会议论文集，第 139 卷，《机器学习研究论文集》，第 10347–10357 页，2021年7月18–24日。[66] C. Wang 和 W. Deng。用于伪造人脸检测的代表性伪造样本挖掘。载于 2021 年 IEEE/CVF计算机视觉与模式识别会议（CVPR），第 14918–14927 页，2021。[67] H. Wang。图像生成中生成对抗网络与扩散模型的比较分析。AMCCE 2024，《计算机科学与智能系统研究汇刊》，120:59–68，2024。[68] J. Wang、K. Sun、T. Cheng、B. Jiang、C. Deng、Y. Zhao、D. Liu、Y. Mu、M. Tan、X. Wang、W. Liu 和 B. Xiao。用于视觉识别的深度高分辨率表征学习，2020。[69] 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 IEEE/CVF计算机视觉与模式识别会议论文集

tion，第8695–8704页，2020年。[70] Y. Wang、D. Yang、F. Bremond 和 A. Dantcheva。潜在图像动画器：通过潜在空间导航学习为图像制作动画。载于国际学习表征会议，2022年。[71] Z. Wang、Z. Chi、Y. Zhang 等人。Fregan：利用频率分量在有限数据下训练生成对抗网络。载于神经信息处理系统进展（*NeurIPS*），第35卷，第33387–33399页，2022年。Yang 等人（2024）引用的官方结果。[72] H. Xie、H. He、B. Fu 和 V. Sanchez。Grdt：利用几何表征分布和纹理实现稳健的深度伪造检测。载于*2025*年*IEEE/CVF*计算机视觉应用冬季会议研讨会（*WACVW*），第686–696页，2025年。[73] C. Xu、J. Zhang、M. Hua、Q. He、Z. Yi 和 Y. Liu。区域感知人脸替换。载于*IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*），第7632–7641页，2022年6月。[74] Z. Yan、T. Yao、S. Chen、Y. Zhao、X. Fu、J. Zhu、D. Luo、L. Yuan、C. Wang、S. Ding 等人。Df40：迈向新一代深度伪造检测。*arXiv*预印本 *arXiv:2406.13495*，2024年。[75] Z. Yan、Y. Zhang、Y. Fan 和 B. Wu。Ucf：揭示可泛化深度伪造检测的共同特征，2023年。[76] Z. Yan、Y. Zhang、X. Yuan、S. Lyu 和 B. Wu。Deepfakebench：深度伪造检测综合基准。载于 A. Oh、T. Neumann、A. Globerson、K. Saenko、M. Hardt 和 S. Levine 编，神经信息处理系统进展，第36卷，第4534–4565页。Curran Associates, Inc.，2023年。[77] Q. Yang、Z. Zhao、Y. Pu、S. Pan、J. Gu 和 D. Xu。Fncontra：对比学习中的频域负样本挖掘，用于有限数据图像生成。《专家系统与应用》，245：123456，2024年。在线发布。[78] X. Yang、Y. Li 和 S. Lyu。利用不一致的头部姿态揭露深度伪造，2018年。[79] L. Yuan、Y. Chen、T. Wang、W. Yu、Y. Shi、Z.-H. Jiang、F. E. Tay、J. Feng 和 S. Yan。Tokens-to-token ViT：在 ImageNet 上从头训练视觉变换器。载于*IEEE/CVF* 国际计算机视觉会议论文集（*ICCV*），第558–567页，2021年10月。[80] W. Zhang、X. Cun、X. Wang、Y. Zhang、X. Shen、Y. Guo、Y. Shan 和 F. Wang。SadTalker：学习逼真的3D运动系数，用于风格化音频驱动单图说话人脸动画。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第8652–8661页，2023年。[81] Z. Zheng、X. Peng、T. Yang、C. Shen、S. Li、H. Liu、Y. Zhou、T. Li 和 Y. You。Open-Sora：让所有人都能使用高效视频制作，2024年。[82] M. Zou、B. Yu、Y. Zhan、S. Lyu 和 K. Ma。面向深度伪造检测的语义导向多任务学习：一种联合嵌入方法，2024年。



ShukeshReddy目前是 BITS Pilani海得拉巴校区计算机科学与信息系统系的博士生。他于 2023 年获得计算机科学学士学位。在攻读博士学位之前，他曾担任软件工程师，负责领导物联网平台和物联网开发套件的开发。他的研究兴趣集中在人脸表征学习，目前的工作致力于推进人脸伪造检测技术。



Srijan Das 我现任北卡罗来纳大学夏洛特分校计算机科学系助理教授。在北卡罗来纳大学夏洛特分校，我的研究方向包括视频表征学习和机器人视觉。我是 AI4Health中心的成员，也是北卡罗来纳大学夏洛特分校夏洛特机器学习实验室（CharMLab）的创始成员之一。



Abhijit Das 现任 BITS Pilani海得拉巴校区助理教授。此前，他曾在法国 Inria Sophia Antipolis–M´éditerran´ée 担任博士后研究员。他在澳大利亚格里菲斯大学信息与通信技术学院获得博士学位。他是一位杰出的机器学习和计算机视觉研究人员，拥有 15 年以上的研究与教学经验。目前，他正在开展一项关于表征学习和人体分析的研究，采用

基于面部和身体的视觉特征。