配对采样对比框架用于联合物理数字人脸攻击检测

Andrei Balykin*
IDRND
andrew.balykin@idrnd.net

Anvar Ganiev*
IDRND
anvar.ganiev@idrnd.net

Denis Kondranin*
IDRND
denis.kondranin@idrnd.net

Kirill Polevoda
IDRND
kirill.polevoda@idrnd.net

Nikolai Liudkevich
IDRND
lyudkevich@idrnd.net

Artem Petrov
IDRND
petrov@idrnd.net

Abstract

现代人脸识别系统仍然容易受到欺骗尝试的影响，包括物理展示攻击和数字伪造。传统上，这两种攻击方式通常由不同的模型或管道来处理，每个都针对其特定的特征和模式。然而，维护不同的检测器会导致系统复杂性增加、推理延迟更高以及综合攻击向量的问题。我们提出了一种配对采样对比框架，这是一种统一的训练方法，利用自动匹配的真实与攻击自拍照对来学习无关模式的生命迹象。在第六届面部防欺骗挑战“统一物理-数字攻击检测”基准测试中，我们的方法获得了 2.10%的平均分类错误率（ACER），优于先前的解决方案。所提出的框架是轻量级的，仅需 4.46 GFLOPs，并且训练运行时间不到一小时，使其适合实际部署。代码和预训练模型可在 https://github.com/xPONYx/iccv2025_deepfake_challenge 获取。

1. 介绍

电子 KYC（了解你的客户）系统在现代安全和认证应用中无处不在，从移动设备解锁到金融交易。尽管准确性与速度方面取得了显著进展，这些系统仍然容易受到广泛类型的攻击。一方面，展示攻击，其中对手向传感器展示诸如打印照片、重放视频或 3D 面具等物理物品，可以通过利用表面视觉线索来规避存活性检查。另一方面，数字攻击利用生成模型在像素级别创建高度逼真的面部伪造物，使身份互换、面部表情动画、口型同步以及其他难以被专家观察者察觉的操纵成为可能。

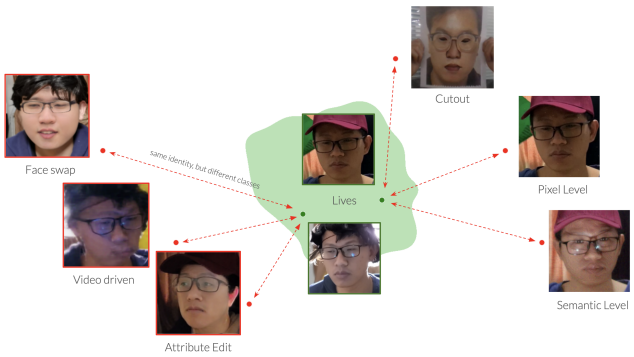


图 1. 所提出框架的核心是从数据集中抽取具有相同用户身份的实时攻击样本对，这对于抵消由于真实样本数量有限而引入的偏差至关重要。

由于两种类型的攻击都在不断发展，实物欺骗的质量提高以及深度伪造生成器产生的视觉伪影减少，仅仅将呈现攻击检测和深度伪造检测作为独立任务处理已不再足够。统一的防御必须准确地检测任何企图愚弄面部识别系统的尝试，无论是通过镜头前的欺骗还是数字操纵视频。最近的基准测试，如 ICCV 2025 “第六届面部防欺骗：统一物理-数字攻击检测”赛道及其相关的 UniAttackData 数据集 [1]，强调了这一联合问题的重要性和难度。仅在 PAD 数据上训练的模型通常会在数字伪造上失败，反之亦然，由于差异化的线索和领域差距 [2]。

*Equal contribution

在这项工作中，我们提出了一种配对采样对比框架，该框架在单模型架构下统一了 PAD 和 DFD。我们在图 1 中提出的思路是形成真实和攻击人脸嵌入的匹配对，并仅将不对称增强应用于真实样本。然后，一个 ConvNeXt-v2-Tiny 主干同时优化二元分类损失（以分离活体与攻击）和监督对比损失（以拉近真实特征并推开所有攻击特征），而 CutMix 正则化模拟部分遮挡，并鼓励网络依赖分布式的活性提示。挑战结果表明，我们的方法实现了平均分类误差率 (ACER) 2.10%，排名在顶尖解决方案之中。

2. 相关工作

基于图像的物理展示攻击和数字伪造检测方法通过分析单个图像来识别欺诈行为，而不依赖于时间线索。最近的进步涵盖了多样的模型架构、泛化策略和基准数据集。

模型架构。基于 CNN 的方法，特别是 EfficientNets [3]，由于其强大的特征提取能力和对细微视觉伪影的鲁棒性而受到欢迎 [4]。然而，CNN 模型在泛化到训练数据集之外时往往面临局限性。视觉变换器 (ViTs)[5] 近来作为有效的替代方案出现，利用图像内的长距离依赖关系来检测操作不一致性。例如，Luo 等人介绍了在预训练的 ViTs 上引入自适应模块，显著提高了跨数据集检测精度 [6]。多任务模型如 LAA-Net 结合了 CNN 主干网络与注意力机制，明确专注于局部混合伪影以识别高质量的深度伪造视频 [7]。此外，诸如 TruFor 这样的混合框架通过变换器架构融合视觉特征和噪声残差，提高了对多种篡改类型的检测能力 [8]。

泛化策略。核心挑战仍然是跨不同深度伪造数据集的泛化。数据增强技术如压缩、模糊和色彩调整有效提升了模型的鲁棒性 [9]。在 [10] 中，作者通过混合真实人脸生成合成数据，展示了改进的泛化能力，创建了类似伪造的伪影而无需特定于深度伪造的模型。身份不变学习也起到了关键作用，通过减少隐式身份偏见，确保模型关注的是操纵伪影而不是特定面部身份 [11]。

数据集。标准化基准驱动检测方法的评估。FaceForensics++ 广泛用于初始训练，但其局限性需要在 Celeb-DF 和 DFDC 等更具挑战性的数据集上进行测试以评估泛化能力 [12–14]。其他数据集如 WildDeepfake 提供现实世界的复杂情况，进一步挑战模型超越受控场景的有效泛化能力 [15]。

联合物理和数字攻击数据集。新出现的威胁扩大了深度伪造检测研究的范围，同时应对数字操作和物理呈现攻击（PAD）。这些综合数据集整合了多样化的攻击向量，提供了更丰富的基准测试，检验探测器在各种攻击模式下的鲁棒性。GrandFake [16] 合并了 PAD 和数字深度伪造数据集，涵盖了 25 种不同的攻击类型。这一全面的数据集有助于开发统一的检测方法，能够同时识别多种伪造和欺骗尝试。同样，在 [2] 中，合并了九个不同的 PAD 和伪造数据集，强调了身份混淆的关键问题。它促进了对探测器易受身份特定偏见影响的分析，并支持旨在创建身份不变检测模型的战略。UniAttackData [1] 通过在物理和数字攻击中保持每个主体的身份一致性，实现了显著的进步。这一设计独特地使模型能够学习一致的身份特征，从而提高对各种类型操作的一般化和鲁棒性。

总结来说，基于图像的物理展示攻击和数字伪造检测方法通过整合先进的架构和泛化策略有了显著的发展。持续的进步依赖于结合这些创新以保持对抗日益复杂的伪造行为的强大能力。

3. 方法论

在本节中，我们概述了所提出方法的三个核心组件。3.1 节介绍了我们的数据过滤管道。随后，3.2 节包含了训练期间使用的配对设置的详细信息。最后，3.3 节解释了用于表示学习的对比训练过程。

3.1. 数据过滤

在对训练数据集进行手动分析时，我们遇到了大量不包含人脸或任何有价值信息的实时样本。为了确保训练设置中仅包括有效的实时样本，我们手动过滤了实时子集，从原始集合中删除了大约 10%的无效图像。

为了消除仅出现在欺骗子集中的身份可能带来的偏差，我们提出了一种基于面部识别系统的过滤流水线。首先，我们使用预训练的 InceptionResNet-v1 模型 [17] 为所有训练图像提取 512 维的面部嵌入。同时，我们使用 MTCNN 检测器 [18] 获得面部边界框。

接下来，我们计算每个欺骗攻击样本与训练数据中所有真实样本之间的余弦相似度：

$$\mathrm{sim}(a, b) = \frac{\langle a,\, b \rangle}{\|a\|_2 \, \|b\|_2},$$

其中，$a, b \in \mathbb{R}^{512}$ 分别是 spoof-攻击图像和真实图像的 InceptionResNet-v1 面部嵌入。

对于每个伪造攻击样本，我们找到与其相似度最高的真实样本：

$$b^{\star}(a) = \arg\max_{l \in \mathcal{L}} \operatorname{sim}\big(a, l\big),$$

其中 $\mathcal{L}$ 表示训练集中所有的真实嵌入。

为了进行过滤，我们只保留那些与最相似的真实样本的相似度超过预定义阈值的欺骗攻击样本：

$$\mathcal{D}_{\text{train}} = \big\{ \, (a,\, b^{\star}(a)) \;\big|\; a \in \mathcal{A},\; \operatorname{sim}\big(a,\, b^{\star}(a)\big) > \tau_{sim} \big\},$$

其中 $\mathcal{A}$ 是所有攻击嵌入的集合而 τ_{sim} 是余弦相似度阈值。我们将相似性视为一个可以基于验证子集上的目标指标进行优化的训练超参数。

作为过滤程序（表 1）的结果，由于缺乏足够相似的活体对照，某些类型的展示攻击被完全从训练集中移除。特别是，所有 PAD 攻击如重放、打印和剪切都被彻底过滤掉。此外，数字攻击类型如面部交换和属性编辑几乎全部被消除，在基于相似性的过滤后仅剩下一小部分样本。

分类	过滤前	过滤后
Pixel-Level	8 364	5 024
Semantic-Level	3 757	2 349
Video-Driven	1 540	520
Face-Swap	6 160	20
Attribute-Edit	1 476	5
Replay	109	–
Cutouts	79	–
Print	43	–
Live Face	839	751

表 1. 攻击类别中训练子集样本分布的过滤前与过滤后比较。

3.2. 采样设置

受到大量像素级、对抗性和属性编辑攻击的启发，这些攻击在保留原始身份的同时引入微妙的操作，我们旨在提高模型对这些细微缺陷的敏感度。为了在一个高度受限的真实样本集上稳定训练，并增加对区分此类缺陷的关注，我们在每个训练批次中利用真实图像和攻击图像的配对采样。

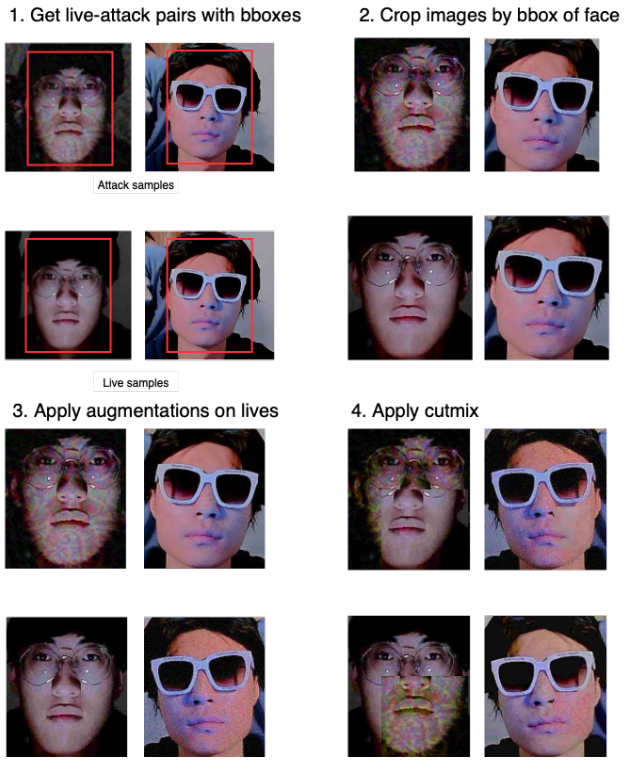


图 2. 我们提出的流水线中用于获得最终训练配对的数据预处理步骤概述：(1) 将攻击与相应的活体进行配对。(2) 使用边界框裁剪图像。(3) 对活体样本应用数据增强。(4) 应用 CutMix 增强策略。

我们利用之前获得的每个攻击样本与其最相似的真实样本之间的映射关系来实施我们提出的方法。首先，我们在批处理中抽取一半的欺骗攻击样本，然后我们将批处理中的每个攻击样本对应的最相似的真实样本附加到这些样本后面。最终训练对集合包含 7 918 对。

令 $\mathcal{S}$ 表示训练数据集中的欺骗攻击样本集合，$\mathcal{L}$ 表示真实样本集合。如果欺骗子集的基数显著大于真实子集的基数，即，

$$|\mathcal{S}| \gg |\mathcal{L}|,$$

则提出的配对策略可以作为真实子集的隐式过采样机制。由于为每个欺骗样本选择最相似的真实样本，来自较小群体 $\mathcal{L}$ 的真实样本更有可能被重复抽样，从而在训练过程中增加它们的影响。

3.3. 对比训练

为了提高所学表示的判别能力，我们优化了一个包含焦点损失 [19] 和监督对比 SupCon 损失 [20] 的双

重目标。令 $\boldsymbol{x}$ 表示输入图像，$\boldsymbol{f}_\theta(\boldsymbol{x}) \in \mathbb{R}^d$ 表示骨干特征向量，而 $\boldsymbol{z} = \mathrm{proj}\big(\boldsymbol{f}_\theta(\boldsymbol{x})\big) \in \mathbb{S}^{d-1}$ 是其在单位球上的 ℓ_2 归一化投影。对于一个 mini-batch$\mathcal{B} = \{(\boldsymbol{z}_i, y_i)\}_{i=1}^N$，锚点 i 的 SupCon 项为：

$$\mathcal{L}_{\text{supcon}}^{(i)} = -\log \frac{\sum_{j \neq i} \mathbf{1}[y_j = y_i] \exp\left(\frac{\boldsymbol{z}_i^\top \boldsymbol{z}_j}{T}\right)}{\sum_{k \neq i} \exp\left(\frac{\boldsymbol{z}_i^\top \boldsymbol{z}_k}{T}\right)} \tag{1}$$

$$\mathcal{L}_{\text{supcon}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{supcon}}^{(i)}, \tag{2}$$

其中 T 是温度超参数。

4. 实验

4.1. 实验设置

数据集和协议。所有实验均在第六届面部防伪挑战赛——统一物理数字攻击检测@ICCV 2025（FAS-ICCV 2025）发布的训练和开发包上进行。该数据集包含 23 367 **张图像**，只有 839 张图像是真实的，而剩下的 21 528 表示来自九种不同攻击类别的伪造样本。这些包括展示攻击（打印、重放、剪切）和数字篡改（属性编辑、换脸、视频驱动合成以及像素级和语义级对抗扰动）。

数据增强。为了缓解类别不平衡并提高泛化能力，我们在训练期间应用了一个基于 Albumentations 的数据增强管道：

- 几何变换：水平翻转，
- 光度测量：亮度/对比度（±10%），色相/饱和度（±10），伽玛（范围 80 – 120），
- 压缩：JPEG 质量从 [40, 60] 中采样。

此外，我们以概率 =0.3 应用了 CutMix 增强，并对真实和攻击样本都使用了 α=0.6，这鼓励模型定位特定于欺骗的伪迹并减轻过拟合。

网络架构。我们的模型骨干是 ConvNeXt-v2-**微型**(28M 参数)[21]，在 ImageNet1K 上预训练 [22]。每个输入人脸裁剪图像被调整为 $224 \times 224 \times 3$(标准 RGB)，由骨干网络全局平均池化，然后传递到两个头部：

1. 一个用于活体/攻击预测的全连接层二分类头
2. 一个投影头，两层 MLP 用于生成对比训练的 128 维嵌入。

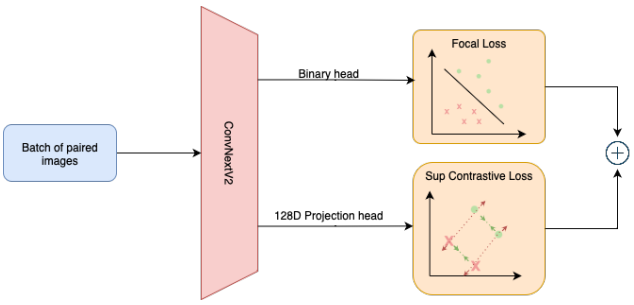


图 3. 我们训练策略中所采用的损失函数组合。

损失函数。训练最小化焦损和监督对比损失的加权和：

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda \mathcal{L}_{\text{supcon}}$$

训练目标结合了带有类别平衡因子 $\alpha = 0.5$ 和聚焦参数 $\gamma = 0.7$ 的焦损，以应对标签不平衡，并在整个身份一致的活体-攻击对的小批量上评估监督对比损失。

我们仅使用 CutMix 增强的输入及其混合标签来专供 Focal 损失头部，而 SupCon 头部则操作原始的未扰动标签。这种分离防止了由 CutMix 生成的混合标签破坏 SupCon 损失的正样本掩码。

评估指标。 比赛排行榜根据隐藏测试集上的**平均分类误差率** (ACER) 进行排名。为了完整性，我们额外报告

- **等错误率** – 接收者操作特性曲线操作点处的等错误率，其中 FAR = FRR;
- **误接受率** – 全部攻击展示分类错误率
- BPCER – 总体真实展示分类误差率
- Acer – APCER 和 BPCER 值的平均值;

选择验证 EER 最低的检查点作为最终提交。

指标定义为：

$$\text{EER:} \quad \text{FAR} = \text{FRR},$$

$$\text{ACER} = \frac{\text{APCER} + \text{BPCER}}{2},$$

$$\text{APCER} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

$$\text{BPCER} = \frac{\text{FN}}{\text{FN} + \text{TP}}.$$

训练详情。 我们对调整大小后的面部裁剪进行了骨干网络的微调。对于数据过滤，我们将 $\tau_{sim} = 0.9$ 设为余

Setup	ACER↓	Acc↑	AUC↑
w Lives Augs & SupCon	0.0085	0.9831	0.9910
w/o SupCon	0.0177	0.9646	0.9871
w/o Lives Augmentations	0.0862	0.9046	0.9646

表 2. 不同训练设置下的验证结果。

弦相似度阈值。每批处理 32 张图像，并在 2x NVIDIA L4 24GB GPU 上进行；每个名义上的周期跨越训练集的一半，运行持续 20 个周期。优化采用 AdamW 并带有权重衰减 1.1×10^{-5}。一个 5%的预热余弦计划在一个周期内将速率从 1.82×10^{-4} 衰减到 6.8×10^{-7}。目标结合了二元焦损失（$\alpha = 0.5$，$\gamma = 0.7$）与监督对比损失（投影维度 128，温度 $\tau = 0.14$，权重 $\lambda = 0.3$）。CutMix 以 0.3 的概率应用于活体和攻击作物，并且是 $\alpha = 0.6$。达到最低验证等错误率的检查点被保留用于提交。

4.2. 评估

我们在"The 6th Face Anti-Spoofing: Unified Physical-Digital Attacks Detection@ICCV 2025" 官方验证集上评估了配对活性对比框架。实验分别在有和没有关键组件的情况下进行：数据配对、非对称实时帧增强以及监督对比学习，这些组件各自对 ACER 减少的影响详见表 2。我们提出的方法在挑战的隐藏测试集中获得了 2.10%的 ACER。总的来说，这些结果验证了我们的统一框架在同时检测物理和数字面部攻击方面的有效性和实用性。

4.3. 消融研究

数据过滤阈值. To determine the optimal cosine similarity threshold τ_{sim} for pairing attack samples with their most similar live counterparts, we conducted a grid search over threshold values from 0.84 to 0.91. Lower thresholds admit more attack–live pairs into the training set, increasing dataset size but also allowing a higher proportion of loosely matched identities. Therefore, higher thresholds ensure stronger identity consistency between pairs, but overly restrictive filtering risks removing diverse training examples. As shown in Tab. 3, the threshold $\tau_{\text{sim}} = 0.90$ yielded the lowest ACER (1.08%) and the highest accuracy (97.85%) on the validation set. This indicates that stricter matching leads to more consistent identity pairs, enabling the model to focus on liveness-related cues. Based on these results, we adopted $\tau_{\text{sim}} = 0.90$ for our experiments.

Threshold	Dataset Size	ACER↓	Acc↑	AUC↑
0.84	10339	0.1426	0.8686	0.9550
0.85	10131	0.0661	0.9446	0.9749
0.86	9887	0.0621	0.9527	0.9850
0.87	9637	0.0785	0.9199	0.9673
0.88	9362	0.0789	0.9192	0.9706
0.89	9036	0.0769	0.9231	0.9747
0.90	8669	0.0108	0.9785	0.9903
0.91	8207	0.0594	0.9581	0.9908

表 3. 余弦相似度阈值对配对采样过滤阶段验证性能的影响。

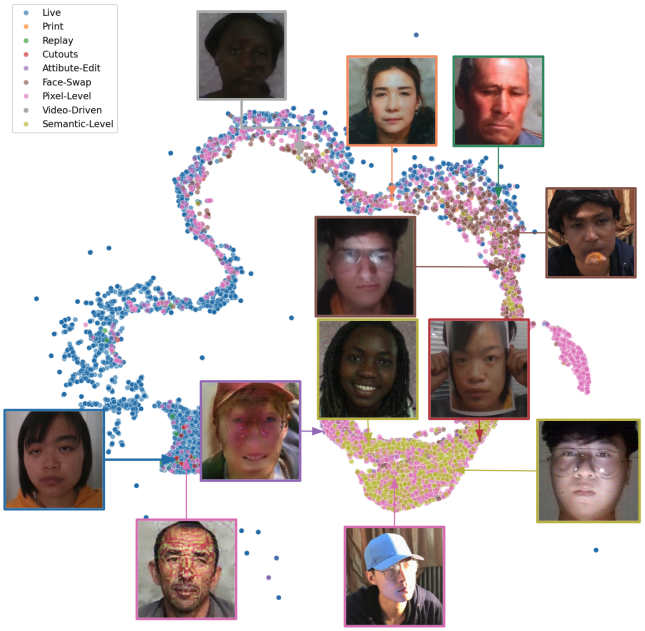
超参数敏感性分析。 为了评估关键增强和对比学习参数对验证性能的影响，我们对 CutMix 参数 α 和 p 以及监督对比损失权重 λ_{supcon} 和温度 T 进行了网格搜索。图 5 可视化了这些超参数作为验证 EER 的函数。

上图显示，低到中等的 CutMix 概率 $p \approx 0.15$–0.3 与 $\alpha \approx 0.6$–1.0 结合倾向于最小化 EER，表明过度使用 CutMix 或过于激进的混合可能会遮蔽活体提示。在底部，最佳的监督对比配置位于 $\lambda_{\text{supcon}} \approx 0.2$–$0.3$ 和 $T \approx 0.15$–0.2 附近，在真实样本的特征紧凑性与攻击嵌入之间的充分分离之间取得平衡。这些结果指导了我们在挑战赛提交中使用的最终超参数设置。

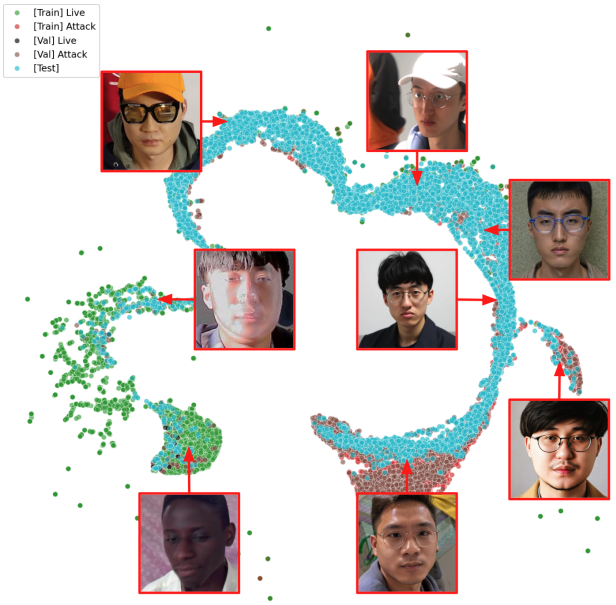
为了更好地理解实时和攻击样本的分布情况，我们对学习到的特征表示进行了 UMAP 分析。UMAP **分析**.

图 4a 展示了各种欺骗类型的 UMAP 投影。每个点对应数据集中的一个样本，颜色编码根据操作类型而定。

我们观察到活体样本的聚类，这些样本与大多数攻击类型相分离。此外，某些欺骗类别倾向于形成局部聚类，表明模型不仅在活体和伪造样本之间学习了判别模式，还在不同类型的攻击中也学到了这种模式。这表明一些攻击类型在其所学特征空间中具有相似的视觉或统计特性。示例图像被叠加以提供这些聚类的可视化背景。



(a) 不同类别的学习特征空间



(b) UMAP 投影按数据集分割分组。

图 4. UMAP 投影用于消融研究。示例图像显示了每个聚类中存在的攻击类型。

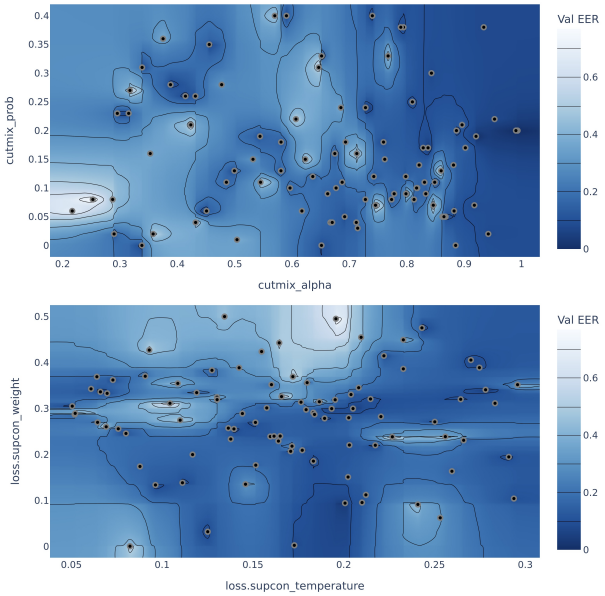


图 5. CutMix 参数（**顶部**）和监督对比损失参数（**底部**）对验证 EER 的影响。较暗的区域表示较低的 EER 值。CutMix 图变化 α 和概率 p，而 SupCon 图变化损失权重 λ_{supcon} 和温度 T。

图 4b 展示了 UMAP 投影图，其中样本根据数据集分割和标签进行分组。我们观察到训练集中的真实样本与验证集和测试集中的明显攻击样本之间存在显著重叠。这表明尽管从人类感知的角度来看存在明显的差异，但学习到的表示有时难以划分清晰边界，可能是由于训练数据的变化有限或在不同分布间泛化的复杂性所致。它们还突出显示了哪些 spoof 类型可能对所提出的方法的泛化更具挑战性。

模型提示. On UniAttackData our model mostly attends to cues that are also obvious to a human observer. As shown in figure 6, (b) and (c) highlight specular reflections and texture inconsistencies on the glasses/skin, while (a) focuses on the mask boundary and local edge discontinuities.

5. 结论

在本文中，我们提出了一种用于面部攻击检测的配对采样对比框架，该框架在一个模型中同时解决了物理展示攻击和数字伪造问题。我们的对比并行训练策略利用了自动配对的真实图像和欺骗图像，并通过结合焦损失和对比损失进行训练以学习鲁棒的生活表

(a) (b) (c)

图 6. 模型注意力集中在攻击上。列车检测器关注人类可见的线索：口罩边缘不连续（1），镜面高光（b）和皮肤纹理异常（c）。

示。我们使用轻量级的 ConvNeXt-v2-Tiny 主干网络与 CutMix 技术实现了这种方法，在 “The 6th Face Anti-Spoofing: Unified Physical-Digital Attacks Detection@ICCV 2025” 挑战中达到了 2.10%的 ACER，使其跻身于表现最佳的方法之列。

广泛的实验表明，我们的方法在挑战数据集上实现了 2.10%的高检测准确率 ACER，同时保持了较低的计算开销（4.46 GFLOPs），适合实时部署。此外，我们的分析强调了对比学习和非对称增强在提高对未见攻击类型和领域的泛化能力方面的益处。

我们相信，一个结合了表现和数字攻击线索的统一检测系统，对于保护下一代人脸识别系统免受不断演变的威胁至关重要。

参考文献

[1] Haocheng Yuan, Ajian Liu, Junze Zheng, Jun Wan, Jiankang Deng, Sergio Escalera, Hugo Jair Escalante, Isabelle Guyon, and Zhen Lei. Unified physical-digital attack detection challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 919–929, 2024. 1, 2

[2] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. IEEE Transactions on Dependable and Secure Computing, 21(5):4327–4342, 2024. 1, 2

[3] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pages 6105–6114. PMLR, 2019. 2

[4] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE transactions on information forensics and security, 15: 42–55, 2019. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 2

[6] Anwei Luo, Rizhao Cai, Chenqi Kong, Yakun Ju, Xiangui Kang, Jiwu Huang, and Alex C Kot Life. Forgery-aware adaptive learning with vision transformer for generalized face forgery detection. IEEE Transactions on Circuits and Systems for Video Technology, 2024. 2

[7] Van Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 17395–17405, 2024. 2

[8] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 20606–20615, 2023. 2

[9] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3994–4004, 2023. 2

[10] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 18720–18729, 2022. 2

[11] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In Proceedings of the IEEE/CVF conference on computer

vision and pattern recognition, pages 4490–4499, 2023. 2

[12] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1–11, 2019. 2

[13] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3207–3216, 2020.

[14] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397, 2020. 2

[15] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM international conference on multimedia, pages 2382–2390, 2020. 2

[16] Debayan Deb, Xiaoming Liu, and Anil K Jain. Unified detection of digital and physical face attacks. In 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pages 1–8. IEEE, 2023. 2

[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, pages 4278–4284, 2017. URL https://arxiv.org/abs/1602.07261. 2

[18] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016. doi: 10.1109/LSP.2016.2603342. URL https://arxiv.org/abs/1604.02878. 2

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 3

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In Advances in Neural Information Processing Systems, volume 33, pages 18661–18673, 2020. URL https://arxiv.org/abs/2004.11362. 3

[21] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16133–16142, 2023. doi: 10.1109/CVPR52729.2023.01548. 4

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://image-net.org/static_files/papers/imagenet_cvpr09.pdf. 4