基于光流的合作性面部活体检测

Artem Sokolov[1, 2], Mikhail Nikitin[2], Anton Konushin[3, 1]

[1] M. V. Lomonosov Moscow State University, Moscow, Russia
[2] Tevian, Moscow, Russia – (artem.sokolov, mikhail.nikitin)@tevian.ru
[3] AIRI, Moscow, Russia – konushin@airi.net

Keywords: Face Liveness Detection, Cooperative Liveness detection, Video Liveness, Face Anti-spoofing, Optical Flow.

Abstract

In this work, we proposed a novel cooperative video-based face liveness detection method based on a new user interaction scenario where participants are instructed to slowly move their frontal-oriented face closer to the camera. This controlled approaching-face protocol, combined with optical flow analysis, represents the core innovation of our approach. By designing a system where users follow this specific movement pattern, we enable robust extraction of facial volume information through neural optical flow estimation, significantly improving discrimination between genuine faces and various presentation attacks (including printed photos, screen displays, masks, and video replays). Our method processes both the predicted optical flows and RGB frames through a neural classifier, effectively leveraging spatial-temporal features for more reliable liveness detection compared to passive methods.

1. 介绍

人脸识别（FR）算法被用于解决需要进行身份识别的广泛实际任务，例如面部支付和通行点控制。作为面对可能处理入侵者的身份识别系统的一部分，使得 FR 算法成为表示攻击的对象。问题在于 FR 模型对待真实和伪造的面孔（可能显示在纸张或屏幕上）是相同的。由于这一特性，如果识别系统没有保护免受此类攻击，任何人的存在就可以被轻易模拟。为了防止在生物识别身份识别过程中身份替换，使用人脸活体检测算法来确定图像中的人脸是真实的还是伪造的。

基本的人脸活体检测方法使用单张输入图像（单次拍摄）进行工作，但在使用高质量假脸时，它们并不总能准确识别对识别系统的攻击。为了提高欺骗分类的准确性，可以利用以视频帧为输入的算法（多次拍摄）。这种方法不仅可以考虑关于面部纹理和周围环境的静态信息，还可以考虑面部运动的时间信息。为了进一步提高准确性，可以在协作模式下录制输入视频，即在拍摄的同时要求用户遵循一组指令。在数据收集过程中施加这种限制可能会增加人员识别的总时间，但会显著降低系统遭受成功攻击的可能性。

在这项工作中，我们提出了一种新颖的合作式多次拍摄面部活体检测方法。算法输入的视频应按照特定场景进行拍摄，其中一个人在两个预定义检查点之间慢慢将面部从正面接近摄像头（“接近面部”场景）。光流检测器被用作视频序列预处理的主要组成部分。所提出方法的新颖之处在于使用了“接近面部”场景拍摄的输入视频的预测面部光流，以进一步进行真实性/伪造性分类。选择的视频场景与光流的结合被证明特别有效，因为这种组合提供了提取面部体积信息的机会。此外，预计的光流与视频中的 RGB 帧一起传递给分类器，以考虑静态和时间信息。

2. 相关工作

视频活体检测领域有着悠久的研究历史。一些早期的方法集中于分析特定的面部运动，例如眼球运动和眨眼模式 (Jee et al., 2006) (Nikitin et al., 2019) 或嘴唇运动 (Kollreider et al., 2007) 。尽管这些方法对于简单的欺骗尝试有效，但它们在针对使用在目标区域预剪开的面具进行的更复杂攻击时却显得脆弱。

另一组视频真实性分类方法是基于现有的单次拍摄方法在视频场景中的适应。例如，(Liang et al., 2024) 引入了一个质量评估模块，使得可以在帧间对单次拍摄模型的预测进行加权，对单次拍摄模型容易分类的帧的预测给予更高的重要性。然而，这种方法由于独立地处理帧，因此无法充分利用时间信息。

视觉变压器的出现使视频分析的新架构成为可能。像 (Ming et al., 2022) 和 (Marais et al., 2023) 这样的作品采用时间注意机制来捕捉帧间关系。虽然这些方法提供了较高的分类准确性，但其计算成本也相当高。

更高效的方法集中于帧聚合。(Parkin and Grinchuk, 2020) 提出了其中一种算法。在这种方法中，作者结合了光流检测器与秩池化 (Fernando et al., 2017) 从输入视频中提取特征进行分类。另一个算法由 (Muhammad et al., 2022) 提出，在该算法中，使用稳定帧平均计算序列聚合以进一步进行活体分类。

另一种基于视频的活体检测方法是合作的方法，这在文献中相对未被充分研究，大多数方法依赖于简单的互动，例如句子发音 (Kollreider et al., 2007) 或面部表情变化 (Ming et al., 2018) 。这种方法增加了验证的时间，但显著提高了准确性。

我们的方法基于这些基础，同时引入了关键创新。像 (Muhammad et al., 2022) 一样，我们采用了稳定化，但通过神经关键点检测来增强它，以改善对齐。类似于 (Parkin and Grinchuk, 2020) ，我们使用了光流，但特别为我们的独特合作场景优化了它，在这种场景中，控制面部运动提供了在被动方法中无法获得的重要体积信息。此外，我们将光流分析与 RGB 帧处理结合起来，以捕捉动态和静态特征，从而提高在从简单的打印输出到复杂的 3D 面具的多种攻击类型中的稳健性。

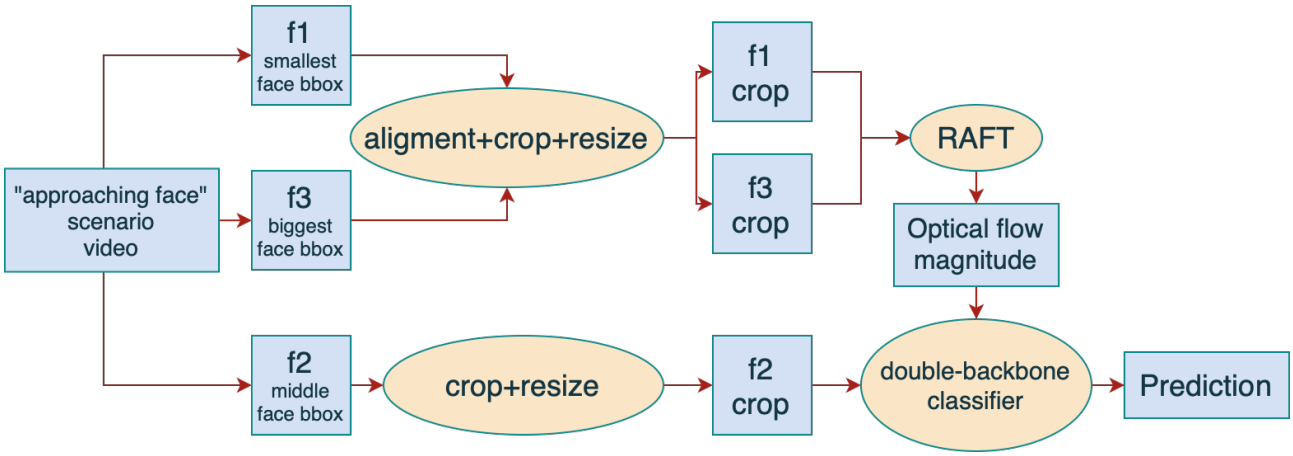

Figure 1. 提出的合作活体检测流程

3. 提出的方法

我们的流程（图 1 ）通过五个阶段收集和处理合作视频：

1. 输入帧捕获：根据合作场景提取第一帧（f1）、中间帧（f2）和最后一帧（f3）（详见第 3.1 节）；
2. 帧预处理：帧 f1、f2 和 f3 被对齐和归一化（详情见 **??** 节）。
3. 光流计算：我们使用预训练的 RAFT 检测器 (Teed and Deng, 2020) 计算 f1 和 f3 之间的光流。
4. 光流处理：通过预测的光流计算幅度，然后进行裁剪（详见第 3.2 节）。
5. 分类：处理过的光流和 RGB 帧 f2 被输入到神经分类器中进行最终预测（详情见第 3.3 节）。

3.1 协作场景

该系统通过在显示屏上可视化录制视频引导用户完成标准化录制协议（图 2 ）。一个实时人脸检测器显示当前的人脸边界框（蓝色方框）。用户首先需要将人脸对齐到一个红色的参考方框（50 % 画面高度），然后慢慢将其移近相机，直到人脸填满一个放大的目标方框（75 % 画面高度）。如果人脸从相机移开或消失，录制将从开始重新进行。这会生成三个关键帧（f1, f2, f3），其相对人脸高度分别为 0.500、0.625 和 0.750，从录制视频中提取出来。

为了使光流检测器能够精准提取视频中脸部大小显著变化的一开始和最后一帧之间的面部运动，我们应用了专门的预处理（图 3 ）：

1. 使用神经检测器检测面部关键点；
2. 通过平移和旋转操作来使 f1 和 f3 帧的关键点居中和对齐；
3. 剪裁面部区域，并在边界框周围留 10 % 的边距；
4. 将所有裁剪调整为 256×256 像素。

的 f2 帧经历相同的处理，除了
对齐步骤。

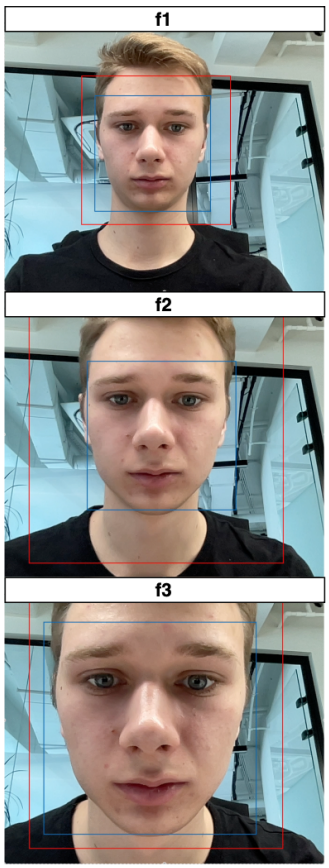

Figure 2. 提出的合作场景

3.2 光流

我们使用预训练的 RAFT (Teed and Deng, 2020) 计算预处理后的 f1 和 f3 帧之间的光流。这揭示了不同攻击类型的独特模式（图 4 ）：

平面静态欺骗（打印照片/屏幕照片）。这种伪造在面部周围的方形区域显示出接近零的流量幅度值（图 4 c-d）。

平面面具。这种伪造在面部区域显示出流量幅度接近零，而在背景上则显示出高值（图 4 ，(b)）。

真实面孔和动态视频回放。光流幅度与面具的相似，但展示了 3D 面部模式以及头部和背景之间更平滑的边缘（图 4 ，(a)）。

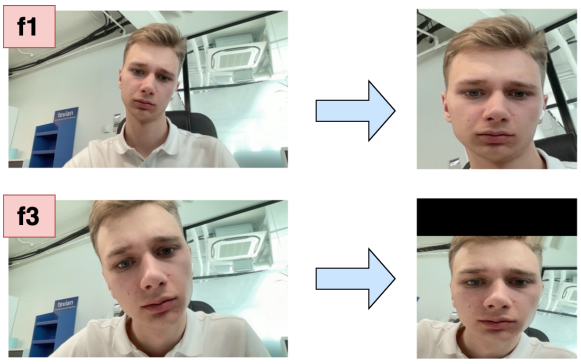

Figure 3. 光流计算前对 f1 和 f3 帧的预处理

由于背景信息没有提供信息量，我们截断超过裁剪尺寸% 20 的幅值（对于 256×256 的输入，51.2 像素）。

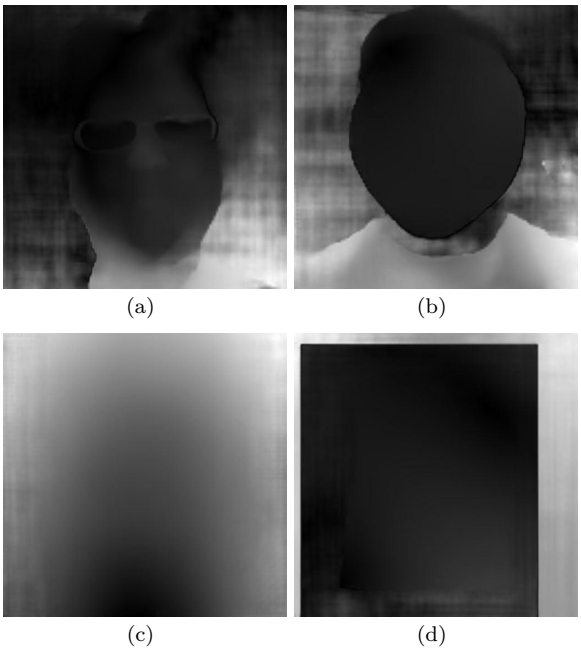

Figure 4. 光流幅度示例：(a) 真人面部 (b) 打印面具攻击 (c) 屏幕照片攻击 (d) 打印照片攻击

3.3 分类模型

为了评估，我们采用了具有全连接融合层的双主干 ResNet18 架构（图 5 ）。第一个“流主干”处理 f1 和 f3 之间截取的光流幅度。第二个“RGB 主干”处理预处理过的 f2 帧。

3.4 训练细节

为了评估所提出的方法，两个骨干网同时进行训练。然而在实际操作中，用于单次活体检测的可用数据量远远多于可以为所选“接近面部”场景收集到的数据量。这就是为什么 RGB 骨干网可以在更大的单次拍摄数据集上单独训练的原因。

为了扩展训练集，我们采用了几种特定于光流的增强方法：

- 随机帧：从面部最小的前 10 个% 帧中随机选择 f1，并从面部最大的前 10 个% 帧中随机选择 f3；

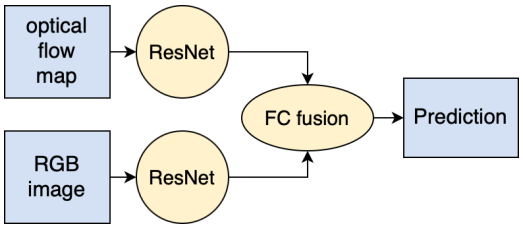

Figure 5. 通过光流图和 RGB 图像进行活体分类的双主干架构

- 多分辨率：计算的流分辨率从 192x192 到 320x320，然后调整为 256×256；
- 透视增强：在预处理之前对 f1 和 f3 帧应用随机透视变换，以提高对头部旋转的鲁棒性。

4. 数据与实验

4.1 数据集

由于我们提出的方法依赖于一种新颖的“接近面部”捕捉场景，现有的公共数据集不适合训练和评估，因为它们不符合这一特定协议。为了解决这一差距，我们收集了一个专门的私人数据集。所有样本都是在各种设备和光照条件下捕获的，记录是由多名参与者提供的。该数据集严格遵循“接近面部”场景，包含真实面部和各种类型的欺骗（如打印照片、屏幕重播）的视频。详细的样本统计信息见表 1 。该数据集包括以下子集：

- 真实：包含真人面孔的图像集合；
- 屏幕照片。子样本拍摄了在不同显示器上展示的人脸照片；
- 打印照片。用打印的照片拍摄的子样本；
- 打印面具。使用打印的全脸面具和剪切面部区域的子样本拍摄；
- 动态视频。子样本是根据“接近面部”场景拍摄的预录视频；
- 静态视频。使用几乎为静态面孔的预录视频拍摄的子样本；

Videos type	train split	test split
Real	1860	291
Screen Photos	4623	217
Printed Photos	1679	85
Printed Masks	1699	93
Dynamic Videos	870	170
Static Videos	765	218

Table 1. 私有数据集统计（样本数量）。

4.2 评估指标

我们使用 ROC AUC 指标来衡量性能，该指标是分别针对每个欺诈子集（类别 0）与真实面部子集（类别 1）计算的。

4.3 RAFT 参数选择

光流检测器的分辨率和细化迭代次数显著影响计算效率和人脸体积信息的提取。由于光流计算是所提出流程的瓶颈，我们尝试通过使用较小分辨率的输入数据和减少细化迭代次数来加快速度。根据图 6 和 7 中所示的实验，接下来的实验中，我们固定使用 256x256 的分辨率和 3 次细化迭代。

所有的时间测量均在 i5-11600K CPU 上的单线程模式下进行。

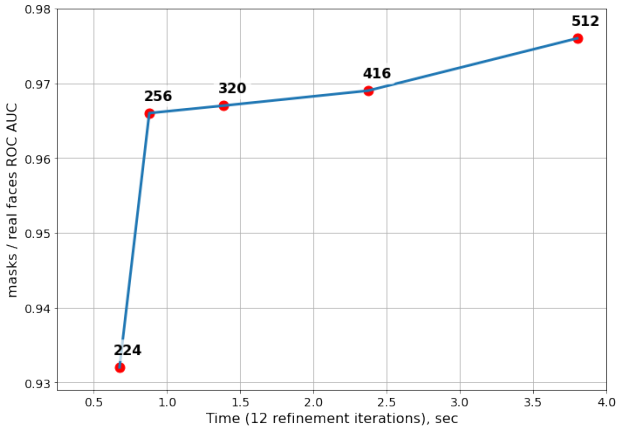

Figure 6. 选择 OF 分辨率

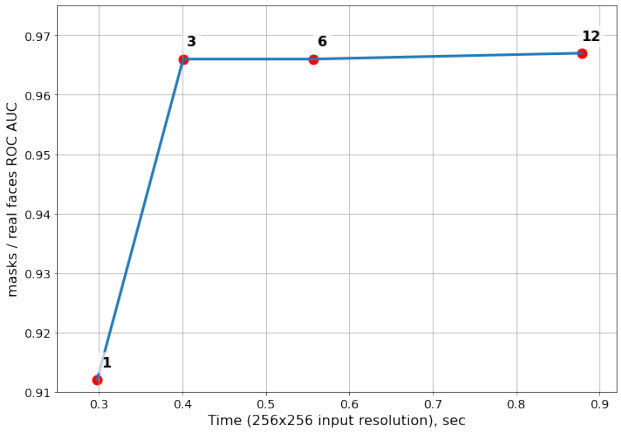

Figure 7. 选择 RAFT 细化迭代次数

4.4 消融研究

作为基准模型，我们使用单一 ResNet18 分类器，该分类器以大小为 256x256 的光流图作为输入。在我们所有的实验中，光流图是通过 RAFT 的 3 次细化迭代计算得到的。

4.4.1 光流处理 光学流处理的几个选项被考虑了：

- 原始光流。具有 X/Y 轴像素位移的双通道输入；
- 光流大小。单通道输入，每个像素的位移距离：$magnitude_{ij} = \sqrt{(x\_shift_{ij} - y\_shift_{ij})^2}$ ，其中 i, j 为一个像素的坐标；
- 剪裁的光流幅度。幅度值如果大于输入裁剪边的% 20，则剪裁为 $inp\_size * 0.2$ 。

表格 2 显示，动态视频子样本的 ROC AUC 显著低于其他样本。这是因为此类欺骗的光流与真实人脸的光流相似。

在后续实验中，我们使用裁剪后的光流幅度作为分类器的输入。

4.4.2 光流增强 为了扩展光流的训练集，我们使用了在 3.4 中讨论的增强方法。表 3 显示，随机帧增强显著促进了 ROC AUC 的提升，而多分辨率方法并没有影响指标。然而，为了整体分类器的稳定性，我们在进一步的实验和最终的流程中使用了这两种增强方法。

4.4.3 RGB 输入 为了使模型从输入帧中提取纹理信息，我们修改了架构，使其能够接收 RGB 图像以及光流。这是我们评估的两种拟议的架构变体：

- OF + RGB（堆叠）：单个 ResNet18。输入是 RGB 图像和光流幅值的连接；
- OF + RGB（分别）：双 ResNet18。第一个主干处理光流幅值，第二个处理 RGB 图像。

表 4 中描述的实验表明，模型提取纹理信息的能力显著提高了所有子集的 ROC AUC 指标，尤其是动态视频部分。

在进一步的实验中，我们使用双 ResNet18 结构，因为它在所有数据子集上表现优于其他模型。

为了增强模型对
在 f1 和 f3 帧之间的头部旋转稳定性，采取了以下措施：

表格 5 显示 ROC AUC 指标不受所讨论变更的影响。这是因为当前的测试分割不包含具有强烈头部旋转的样本。然而，手动测试表明，在将建议的修改添加到最终流水线后，流程对所有类型的头部旋转都变得稳定。

4.4.4 模糊 我们检查了所提出的流水线在模糊图像上的稳定性。我们使用了来自 OpenCV 的高斯模糊函数来创建模糊图像。在我们的实验中，我们使用了三个模糊级别：低（内核大小为平均 f1 脸部裁剪高度的 1 %），中（内核大小为平均 f1 脸部裁剪高度的 6 %）和高（内核大小为平均 f1 脸部裁剪高度的 12 %）（表 6 ）。

4.5 与其他聚合方法的比较

我们还将 (Muhammad et al., 2022) 和 (Parkin and Grinchuk, 2020) 中的想法应用于提出的合作场景，以便与开发的流程进行比较（表格 7 ，8 ）。选择这些方法进行比较是因为我们发现它们在” 接近面部” 场景中使用是有益的。以下是这些方法适应的详细信息：

这个想法借鉴自 (Muhammad et al., 2022) 。为了改进面部稳定性，我们使用了神经网络面部关键点检测器。进行的实验表明，使用 5 帧来计算输入视频聚合可以实现最佳的算法质量。我们使用了 ResNet18 模型进行分类。

这个想法取自 (Parkin and Grinchuk, 2020) 。在改进后的流程中，我们对输入视频应用两次秩池化，参数与原论文中相同。此外，我们在框架 f1 和 f3 之间使用 RAFT 计算光流，如所提议的方法中所示。接收到的特征被传递给三个 ResNet18 骨干网络，并与全连接融合层结合，如所提议的方法中所示。

我们还将提出的管道与在 f2 帧上评估的单次 ResNet18 模型进行了比较。

OF preprocessing type	Screen	Printed	Masks	Dynamic Videos	Static Videos
OF	0.992	0.990	0.953	0.668	0.958
OF magnitude	0.991	0.991	0.955	0.670	0.961
clipped OF magnitude	0.994	0.993	0.965	0.784	0.965

Table 2. 光流预处理的比较（ROC AUC）

Augmentations	Screen	Printed	Masks	Dynamic Videos	Static Videos
no augmentations	0.994	0.993	0.965	0.784	0.965
random frame	0.998	0.997	0.972	0.666	0.972
random frame + multi resolution	0.998	0.996	0.972	0.675	0.972

Table 3. 光流增强的比较（ROC AUC）

Architecture	Screen	Printed	Masks	Dynamic Videos	Static Videos
OF only; ResNet18	0.998	0.996	0.972	0.675	0.972
OF + RGB (stacked); single ResNet18	0.996	0.994	0.985	0.991	0.998
OF + RGB (separately); double ResNet18	0.999	1.000	0.993	0.994	0.999

Table 4. 架构比较（ROC AUC）

Modification	Screen	Printed	Masks	Dynamic Videos	Static Videos
None	0.999	1.000	0.993	0.994	0.999
perspective augmentations	1.000	1.000	0.994	0.996	0.999
perspective augmentations + stabilization	1.000	1.000	0.994	0.996	0.999

Table 5. 增强模型在有头部旋转样本上的稳定性的修改（ROC AUC）

Blurriness level	Screen	Printed	Masks	Dynamic Videos	Static Videos
no blur	1.000	1.000	0.994	0.996	0.999
low	0.997	0.996	0.992	0.989	0.999
medium	0.990	0.981	0.939	0.966	0.984
high	0.937	0.883	0.833	0.846	0.857

Table 6. 在模糊测试上的评估（ROC AUC）

Aggregation	Screen	Printed	Masks	Dynamic Videos	Static Videos
single shot	0.983	0.963	0.970	0.991	0.985
stabilized average (Muhammad et al., 2022)	0.999	0.980	0.922	1.000	1.000
rank pool + OF (Parkin and Grinchuk, 2020)	0.993	0.982	0.904	0.921	0.976
proposed method	1.000	1.000	0.994	0.996	0.999

Table 7. 聚合类型比较（ROC AUC）

Method	sec.
single-shot	0.05
stabilized average	0.05
rank pool + OF	0.75
proposed method	0.55

Table 8. 运行时性能比较（不包括预处理）

5. 结论

在这项工作中，我们提出了一种新颖的视频协作人脸活体检测方法，该方法利用在受控“接近人脸”场景下的光流分析。我们的方法结合了稳定的人脸关键点对齐和神经光流估计，有效地区分真实人脸和包括打印照片、屏幕显示、面具和视频回放在内的各种欺骗攻击。

References

Fernando, B., Gavves, E., Oramas M., J., Ghodrati, A., Tuytelaars, T., 2017. Rank Pooling for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 773–787. http://dx.doi.org/10.1109/TPAMI.2016.2558148.

Jee, H.-K., Jung, S., Yoo, J.-H., 2006. Liveness Detection for Embedded Face Recognition System. *International Journal of Biological and Medical Sciences*, 1, 235-238.

Kollreider, K., Fronthaler, H., Isaac, M., Bigun, J., 2007. Real-Time Face Detection and Motion Analysis With Application in “Liveness” Assessment. *Information Forensics and Security, IEEE Transactions on*, 2, 548 - 558.

Liang, Y.-C., Qiu, M.-X., Lai, S.-H., 2024. Fiqa-fas: Face image quality assessment based face anti-spoofing. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) , 1462–1470.

Marais, M., Brown, D., Connan, J., Boby, A., 2023. Facial liveness and anti-spoofing detection using vision transformers. Proc. Southern Afr. Telecommun. Netw. Appl. Conf.(SATNAC) , 1–6.

Ming, Z., Chazalon, J., Luqman, M. M., Visani, M., Burie, J.-C., 2018. Facelivenet: End-to-end networks combining face verification with interactive facial expression-based liveness detection. 2018 24th In-

ternational Conference on Pattern Recognition (ICPR) , IEEE, 3507–3512.

Ming, Z., Yu, Z., Al-Ghadi, M., Visani, M., MuzzamilLuqman, M., Burie, J.-C., 2022. Vitranspad: Video transformer using convolution and self-attention for face presentation attack detection.

Muhammad, U., Yu, Z., Komulainen, J., 2022. Self-supervised 2d face presentation attack detection via temporal sequence sampling. *Pattern Recognition Letters*, 156, 15–22.

Nikitin, M. Y., Konushin, V. S., Konushin, A. S., 2019. Face anti-spoofing with joint spoofing medium detection and eye blinking analysis. *Computer Optics*, 43(4), 618–626.

Parkin, A., Grinchuk, O., 2020. Creating artificial modalities to solve rgb liveness.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 , Springer, 402–419.