ICS 35.240.40
CCS A11

团　　体　　标　　准

T/ZFIDA 0003—2024

虚假数字人脸检测金融应用技术规范

Technical specification for financial applications of deepfake detection

2024-09-06 发布　　2024-09-06 实施

中关村金融科技产业发展联盟　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由蚂蚁科技集团股份有限公司提出。

本文件由中关村金融科技产业发展联盟归口。

本文件起草单位：萨思数字科技（北京）有限公司、中关村金融科技产业发展联盟、北京国家金融科技认证中心股份有限公司、蚂蚁科技集团股份有限公司、中国工商银行股份有限公司、中国建设银行股份有限公司、中国邮政储蓄银行股份有限公司、中信银行股份有限公司、浙江网商银行股份有限公司、北京前沿金融监管科技研究院、蚂蚁区块链科技（上海）有限公司、中关村互联网金融研究院、北京一砂信息技术有限公司、中移动金融科技有限公司、博彦科技股份有限公司、思旦达（北京）技术有限公司、北京车晓科技有限公司。

本文件起草人：罗曼、林冠辰、朱凯、刘勇、曹亭亭、李振、李博文、陆碧波、彭晋、李哲、王维、李建树、姚伟斌、杨小强、方蕊、赵浚雅、尉郭晨、叶红、程佩哲、许啸、陈天涵、李婧、刘丽娟、康亚冰、史佳涵、赵汉杰、王仲实、张然、张园超、王龙、江嘉航、张克、姜楠、徐睿阳、王丽娜、陈曦、张楚、路如毅、陈明、高丽、刘江涛、谭福铃、陈振、谢源、刘智江。

虚假数字人脸检测金融应用技术规范

1 范围

本文件规定了面向金融领域应用的虚假数字人脸检测服务的技术框架、功能要求、性能要求等，并提出对应的测试评估方法。

本文件适用于金融领域虚假数字人脸检测相关产品研发设计、测试和应用，也适用于第三方评测机构对虚假数字人脸内容检测服务的技术安全评估。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

虚假数字人脸内容 digitally forged face content

利用生成式人工智能技术、计算机视觉技术、数字伪造技术和深度学习模型所生成的虚假的数字人脸图像、视频内容。

3.2

虚假数字人脸内容检测 digitally forged face content detection

鉴定数字图像或视频中人脸内容是否虚假的检测技术。

注：虚假数字人脸内容检测是面向深度伪造、数字化伪造、对抗攻击的人脸安全检测能力，通常基于卷积神经网络、视觉Transformer、计算机视觉算子等深度学习技术提取输入样本的面部特征和攻击线索进行真假数字人脸鉴别。

3.3

虚假数字人脸内容检测服务 service of digitally forged face content detection

提供虚假数字人脸内容检测能力的一体化技术服务。

注：虚假数字人脸内容检测服务是包括检测算法、运行平台、用户系统等软硬一体、交互可控的解决方案等形式。

3.4

准确性 accuracy

虚假数字人脸内容检测服务输出真假人脸检测结果的准确程度。

3.5

鲁棒性 robustness

虚假数字人脸内容检测服务应用到叠加不同干扰的数据内容并做出准确检测的能力。

3.6

泛化性 generalization

虚假数字人脸内容检测服务应用到新攻击方法所扩展的新数据内容并做出准确检测的能力。

4 缩略语

下列缩略语适用于本文件。

AKD：平均关键点距离值（Average Key-point Distance）

GAN：生成对抗网络（generative adversarial network）

HPMSE：头部空间位置差异值（Head Position Mean Square Error）

ID：身份标识（Identity）

PPG：光体积变化描记图法（Photoplethysmography）

PRMSE：旋转角度差异值（Pose Rotation Mean Square Error）

PSNR：峰值信噪比（Peak Signal-to-Noise Ratio）

SSIM：结构化相似性（Structured Similarity）

5 技术框架

人脸识别已经广泛用于金融账户开户、账户登录、移动支付、理财保险身份鉴别等各种金融应用场景。在金融应用场景应用人脸识别时，身份鉴别服务提供者应针对虚假数字人脸进行安全检测，保障相关金融业务安全、顺利开展。

虚假数字人脸检测金融应用技术框架如图1所示，主要包括如下：

a) 功能方面，主要从支持不同伪造类型的虚假数字人脸内容检测能力、 支持包含非图像音频信息的虚假数字人脸内容检测能力、 支持叠加了不同干扰的虚假数字人脸内容检测能力、 支持跨数字人脸内容的联合虚假数字人脸内容检测能力、 支持检测结果的可解释性检测能力等五个维度进行评估；

b) 性能方面，主要通过准确性、鲁棒性、泛化性、响应速度等四个维度的评估指标进行衡量。

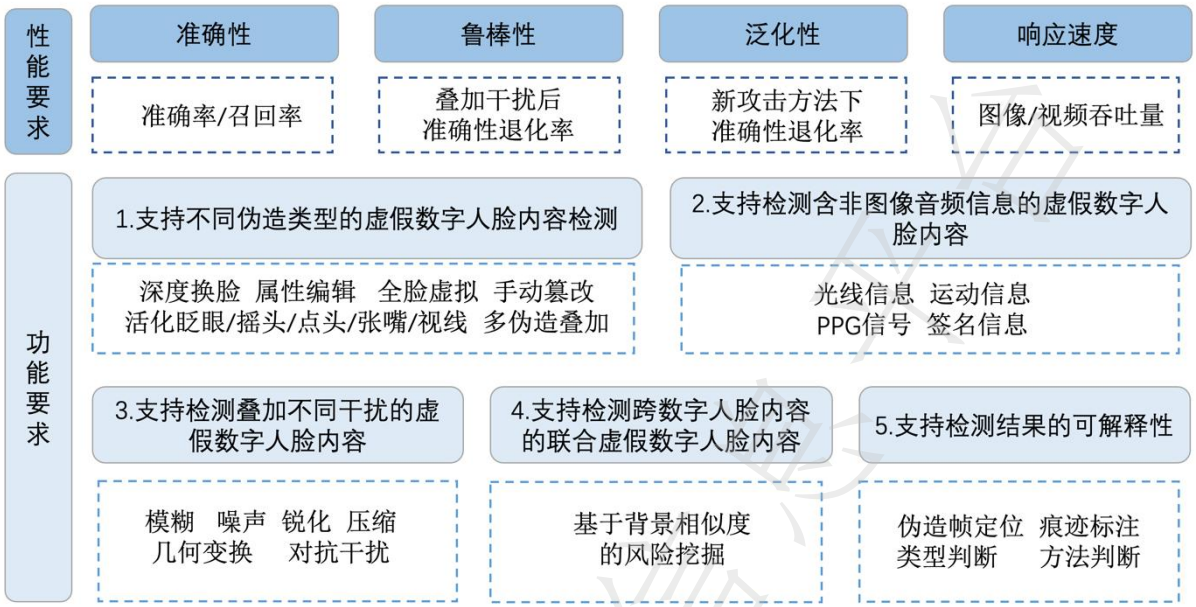


图 1 虚假数字人脸检测金融应用技术框架

6 功能要求与评估

虚假数字人脸检测服务功能要求与评估包括支持不同类型的虚假数字人脸内容检测能力、支持包含非图像音频信息的虚假数字人脸内容检测能力、支持叠加了不同干扰的虚假数字人脸内容检测能力、支持跨数字人脸内容的联合虚假数字人脸内容检测能力、支持检测结果的可解释性检测能力等维度，具体要求和评估指标及查验内容见表1。

表1 功能要求评估指标及查验内容

序号	功能要求	评估方法	详细功能要求	查验内容
1	应支持对不同类型的数字人脸内容进行虚假数字人脸内容检测	使用真实数字人脸数据集，虚假数字人脸数据集测试	评估是否具备对单张图像进行虚假数字人脸内容检测的能力	是否具备对使用了替换方式的单张图像进行虚假数字人脸内容检测的能力
				是否具备对使用了手动篡改方式的单张图像进行虚假数字人脸内容检测的能力
				是否具备对使用了人脸属性编辑方式的单张图像进行虚假数字人脸内容检测的能力

				是否具备对使用了生成式人工智能的单张图像进行虚假数字人脸内容检测的能力
				是否具备对使用了多种伪造技术的单张图像进行虚假数字人脸内容检测的能力
			评估是否具备对无语音的动态视频（包括摇头、眨眼、张嘴、视线变化、手势等等）进行虚假数字人脸内容检测的能力	是否具备对使用了活化技术、生成式人工智能产生的眨眼视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的左右摇头视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的点头视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的张嘴视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的视线变化视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的手势交互视频进行虚假数字人脸内容检测的能力
				是否具备对使用了活化技术、生成式人工智能产生的多种人脸与手势动作的视频进行虚假数字人脸内容检测的能力
			评估是否具备对包含	是否具备对使用了活化技术、生成式人工智能、语音驱动

			语音内容的动态视频进行虚假数字人脸内容检测的能力	技术的音视频进行虚假数字人脸内容检测的能力
2	宜支持包含非图像音频信息的数字人脸内容进行虚假数字人脸内容检测	使用扩展虚假数字人脸数据集测试	评估是否具备对包含从其他传感器获取的光线信息的数字人脸内容进行虚假数字人脸内容检测的能力	
			评估是否具备对包含从其他传感器获取的设备运动信息的数字人脸内容进行虚假数字人脸内容检测的能力	
			是否具备对包含 PPG 信号的数字人脸内容进行虚假数字人脸内容检测的能力	
			评估是否具备对包含从设备获取的签名信息的数字人脸内容进行虚假数字人脸内容检测的能力	
3	宜支持叠加了不同干扰的数字人脸内容进行虚假数字人脸内容检测的能力	使用干扰虚假数字人脸数据集测试	评估是否具备对模糊人脸（人脸像素<1024P）的数字人脸内容进行虚假数字人脸内容检测的能力	
4	宜支持跨数字人脸内容的联合虚假数字人脸内容检测	使用同背景和服饰的真实人脸数据集测试	评估是否具备基于多份数字人脸内容的背景相似度进行虚假数字人脸内容检测的能力	
5	宜支持检测结	现场功能核	可评估是否具备解释	支持伪造帧定位检出

	果的可解释性	验，与虚假数字人脸内容的制作方式进行比较	性检测能力	支持伪造痕迹标注
				支持伪造类型判断
				支持伪造方法判断

7 性能要求与评估

7.1 数据集

7.1.1 真实人脸内容数据集

真实人脸内容数据集构建要求包括但不限于如下：

a) 构建真实人脸内容数据集时，应确保采集的数据具有多维度特征。
b) 数字人脸内容数据应包含不少 5000 张的真实数字人脸内容图像和视频构成的数据集，不应包括计算机生成、篡改、变造的图像和视频数据。
c) 应考虑室内和室外的不同场景，以及绿幕背景的使用，以增加数据集内容的多样性。
d) 数据集应包含从左到右、从右到左、从远到近以及固定镜头模式等多种拍摄角度和视角。
e) 在人物姿态方面，数据集应包括坐姿、站姿和走姿等不同姿势的人物，以展示人物的不同动作和表情。
f) 数据集应包含微笑、大笑、惊悚、愤怒、哭泣、无表情等多种人物表情的多维组合式采集样本。

7.1.2 虚假人脸数据集

虚假数字人脸内容检测数据集构建要求见表2。

注：其中涉及虚假数字人脸视频的数据全部采用相同的每秒帧数（FPS，推荐值为30）。

表2 虚假人脸数据集

数据集名称	指标项	数据集内容要求
基础数据集	对单张图像类虚假数字人脸内容检测的准确性、响应速度	使用人脸替换类技术制作的单张图像不少于 5000 张，其常见形式为将源人物的人脸自动地替换至目标中的人物脸部，使目标中人物身份发生了改变。通常认为身份信息辨识仅与人脸五官特征强相关。如 FaceSwap、HifiFace、DeepFaceLab、FaceShifter、fsgan、DeepFakes、DiffFace 等

		专家篡改类技术制作的单张图像不少于 5000 张，其常见形式为对目标人脸的眼、鼻、嘴、眉、发等五官区域或者人脸外背景区域进行局部抠图、贴图和融合。目标人物主体的身份保持不变，仅发生局部五官或背景的手工篡改操作 常见手工篡改区域包括眼睛、鼻子、嘴巴、眉毛、头发等重点五官区域和人脸以外背景区域等。使用手工篡改技术制作的一张虚假数字人脸图像中的篡改区域数量应从 1 个到 5 个各占比 20%（各不少于 1000 张）
		使用人脸属性编辑技术制作的单张图像不少于 5000 张，其常见形式为改变目标人脸的某些非身份属性特征，比如肤色、发色、年龄、性别、表情、风格、角度、体格、是否佩戴眼镜等等。在人脸属性编辑时，对不同种类的属性进行修改会涉及不同区域和大小的面部变化，并会产生不同程度的伪造线索。因此单张虚假人脸图像中的篡改属性应从 1 个到 5 个各占比 20%（各不少于 1000 张）
		使用人脸生成类技术制作的单张图像不少于 5000 张，如 StyleGAN、WGAN、StarGAN、DCGAN、Stable Diffusion 等
		同时使用了 2 种以上的虚假数字人脸制作技术的单张图像不少于 5000 张
	对无语音的动态视频类虚假数字人脸内容检测的准确性、响应速度	使用人脸活化类技术制作的无语音动态视频不少于 5000 段，其常见形式为使用源人物面部表情或身体姿态驱动目标中人物的表情或姿态。目标人物主体的身份保持不变，仅发生表情或姿态的改变。如 FOMM、Vid2Vid、DaGAN、Wav2Lip、MakeItTalk、X2Face 等
		使用人脸生成类技术制作的无语音动态视频不少于 5000 段，如

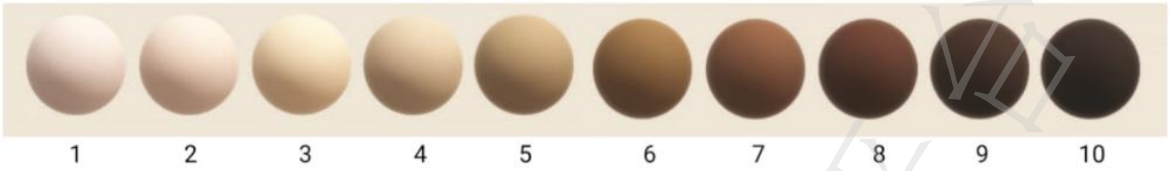
		StyleGAN、WGAN、StarGAN、DCGAN、Stable Diffusion 等。应包括生成不同种族、不同肤色、不同年龄段、不同性别、不同颜值等各种不同类型人脸
		虚假数字人脸进行眨眼的无语音动态视频占无语音的动态视频总量的 20%（不少于 1000 段），应包括全部的制作技术。
		虚假数字人脸进行左右摇头的无语音动态视频占无语音的动态视频总量的 20%（不少于 1000 段），应包括全部的制作技术。
		虚假数字人脸进行点头的无语音动态视频占无语音的动态视频总量的 20%（不少于 1000 段），应包括全部的制作技术。
		虚假数字人脸进行张闭嘴动作的无语音动态视频占无语音的动态视频总量的 20%（不少于 1000 段），应包括全部的制作技术
		虚假数字人脸进行视线变化的无语音动态视频占无语音的动态视频总量的 10%（不少于 500 段），应包括全部的制作技术和左右上下视线
		虚假数字人脸进行手势交互的无语音动态视频占无语音的动态视频总量的 10%（不少于 500 段），应包括全部的制作技术和手指往各个方向挥动
		虚假数字人脸同时进行 2 种以上动作的无语音动态视频占无语音的动态视频总量的 10%（不少于 500 段），应包括全部的制作技术
	对包含语音内容的动态视频类虚假数字人脸内容检测的	使用人脸生成类技术制作的包含语音动态视频不少于 5000 段
		使用人脸活化类技术制作的包含语音动态视频不少于 5000 段

	准确性、响应速度	使用语音驱动技术制作的包含语音动态视频不少于 5000 段
扩展数据集：其中的数字人脸内容包含图像音频外的信息	包含从其他传感器获取的光线信息的数字人脸视频不少于 5000 段，图像不少于 5000 张	
	包含从其他传感器获取的设备运动信息的数字人脸视频不少于 5000 段	
	包含 PPG 信号的数字人脸视频不少于 5000 段	
	包含从设备获取的签名信息的数字人脸视频不少于 5000 段，图像不少于 5000 张	
干扰数据集：其中的数字人脸内容叠加不同程度的干扰	叠加均匀模糊虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加高斯模糊虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加均匀噪声干扰的虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加高斯噪声干扰的虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加锐化干扰的虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加码率压缩干扰的虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加几何变换干扰的虚假数字人脸（人脸像素<1024P）视频不少于 5000 段，图像不少于 5000 张	
	叠加数字对抗干扰的虚假数字人脸（（人脸像素<1024P））视频不少于 5000 段，图像不少于 5000 张	

7.1.3 数字人脸数据集的内容多样性要求

真实人脸数据集和虚假数据集应满足内容多样性的要求，人脸Monk Scale要求见表3。

表3 Monk Scale



应包含不同人员的数字人脸	应包含不同肤色的数字人脸	深肤色（Monk Scale 8~10）数字人脸内容占比至少 30%（不少于 15000 个）
		浅肤色（Monk Scale 1~4）数字人脸内容占比至少 30%（不少于 15000 个）
		中等肤色（Monk Scale 5~7）数字人脸内容占比至少 30%（不少于 15000 个）
	应包含不同种族的数字人脸	撒哈拉以南非洲（非洲中部、东部、西部）数字人脸内容占比至少 10%（不少于 5000 个）
		亚洲东部数字人脸内容占比至少 10%（不少于 5000 个）
		亚洲南部数字人脸内容占比至少 10%（不少于 5000 个）
		亚洲东南部数字人脸内容占比至少 10%（不少于 5000 个）
		欧洲、亚洲西南部、非洲北部、美洲数字人脸内容占比至少 10%（不少于 5000 个）
	应包含不同年龄的数字人脸	0~10 岁的数字人脸占比 10%～20%（5000～10000 个）
		10~18 岁的数字人脸占比 10%～20%（5000～10000 个）
		18~39 岁的数字人脸占比 20%～30%（10000～15000 个）
		39~50 岁的数字人脸占比 20%～30%（10000～15000 个）

		50~65 岁的数字人脸占比 20%～30%（10000～15000 个）
		＞65 岁的数字人脸占比 10%～20%（5000～10000 个）
	应包含脸部遮挡的数字人脸	包含墨镜、眼镜的数字人脸内容占比不少于 10%（不少于 5000 个）
		包含口罩的数字人脸内容占比不少于 10%（不少于 5000 个）
		包含头巾，包巾等特殊种族面部特征的数字人脸内容占比不少于 10%（不少于 5000 个）
		包含脸部彩绘的数字人脸内容占比不少于 10%（不少于 5000 个）
应包含不同环境的数字人脸内容	应包含不同光线环境的数字人脸内容	屏幕主动发光等可控光源环境下的数字人脸内容占比推荐值 10%（不少于 5000 个）
		全黑环境下的虚假数字人脸内容占比推荐值 10%（不少于 5000 个）
		会造成过曝的自然光环境下的数字人脸内容占比推荐值 10%（不少于 5000 个）
		会造成过曝的人工光源环境下的数字人脸内容占比推荐值 10%（不少于 5000 个）
		室内正常环境（200lux）下的数字人脸内容占比推荐值 60%（不少于 30000 个）
	应包含不同背景的数字人脸内容	不发光复杂背景下的数字人脸内容占比推荐值 20%（不少于 10000 个）
		发光复杂背景下的数字人脸内容占比推荐值 20%（不少于 10000 个）
		纯色背景下的数字人脸内容占比推荐值 20%（不少于 10000 个）

		人物背景下的数字人脸内容占比推荐值 20%（不少于 10000 个）
	应包含不同环境音的数字人脸内容	无明显环境音的数字人脸内容占比推荐值 20%（不少于 10000 个）
		环境音音量与目标人物语音大致相同的数字人脸内容占比推荐值 20%（不少于 10000 个）
		环境音音量超过目标人物语音的数字人脸内容占比推荐值 20%（不少于 10000 个）
应包含不同镜头状态的数字人脸内容	应包含晃动镜头状态的数字人脸内容	画面剧烈抖动的数字人脸内容占比不少于 10%（不少于 5000 个）
	应脏污镜头状态的数字人脸内容	模糊、裂纹、脏污等镜头状态的数字人脸内容占比不少于 10%（不少于 5000 个）
应包含人脸面积在整个数字人脸内容中的不同占比	人脸面积 30%～50%的数字人脸内容占数据集总体约 50%（不少于 25000 个）	
	人脸面积大于 50%的数字人脸内容占数据集总体约 50%（不少于 25000 个）	

7.1.4 虚假数字人脸内容数据集质量要求

虚假数字人脸内容数据集应保障如下质量指标具有相同数量的高、中、低指标的虚假数字人脸内容。比如在清晰度维度上，数据集中应包含相同数量的高清晰度、中清晰度、低清晰度的数字人脸内容。同时应保障数据集的平衡性，避免数据集中某些指标的虚假数字人脸过多导致其他指标的虚假数字人脸被忽略。

注：由某些制作技术制作的虚假数字人脸内容，某些维度上不会有明显的变化和差异，比如通过生成式人工智能的虚假数字人脸，其轮廓融合度就比较好，不太会出现轮廓融合度较低的情况。

7.1.5 虚假数字人脸内容整体攻击能力

应使用多种虚假数字人脸内容检测算法对数据集中的内容进行初步测试，评估数据集整体攻击能力的规则如下：

a) 80%的数字人脸内容检测算法检出风险概率分高于 0.8 的虚假数字人脸内容占比 30%；
b) 80%的数字人脸内容检测算法检出风险概率分位于 0.3~0.7 之间的虚假数字人脸内容占比 50%；
c) 80%的数字人脸内容检测算法检出风险概率分低于 0.2 的虚假数字人脸内容占比 20%。

7.1.6 虚假数字人脸内容图像整体质量

应能支持评估虚假数字人脸内容图像整体的质量好坏。可使用如下两个指标对虚假数字人脸内容图像整体质量进行计算：

a) 结构化相似性 SSIM（Structured Similarity）：SSIM 定量评估了图像中亮度、对比度和结构等重要视觉特征，数值越大表示图像整体质量越好。

b) 和峰值信噪比 PSNR（Peak Signal-to-Noise Ratio）：PSNR 定量评估了图像抗压缩失真的能力，数值越大表示图像整体质量越好越稳定。

注：质量越高，就表明整个图像在宏观上的视觉特征更加清晰、色彩信息更加合理、内容结构更加稳定相关，整体也更接近于真实自然的图像信号，相应的攻击能力就会越强。

7.1.7 虚假数字人脸内容图像分辨率

应能支持评估虚假数字人脸内容图像整体的像素分辨率，即宽度和高度的像素数量。

注：像素分辨率通常处于480~2048像素之间。生成的虚假数字人脸内容图像分辨率越大，就意味着图像在单位面积内包含的像素越多，能够展示更多细节内容，从而一定程度上提高图像清晰度，对于虚假数字人脸内容检测服务就有更强的欺骗能力。

7.1.8 虚假数字人脸部质量

应能支持评估虚假数字人脸图像中人脸区域的质量。可使用常见人脸质量打分模型，对人脸的清晰度、完整度、亮度对比度、姿态表情是否自然等进行0~100的综合打分。

注：虚假数字人脸整体的质量越高，就表示人脸区域有着更高的清晰度和完整度，整体姿态表情也是自然合理的，说明脸部的面部畸变、扭曲等线索就更加微弱，相应的攻击能力就会越强。

7.1.9 虚假数字人脸局部质量

应能支持评估对人脸五官（眼、鼻、嘴、眉、发）划分不同局部区域后，对各个局部区域分别评估的质量。可使用常见的五官质量打分算法分别对重点五官进行评估得到虚假数字人脸局部质量。

注：局部五官的细节纹理重建在目前没有被完全解决，尤其是眼睛和嘴部等蕴含丰富特征的区域，对人脸局部五官进行更精细化的质量判断可以确定五官细节特征的真实度。虚假数字人脸中这些重点局部区域的质量分越高，相应的攻击能力就会越强。

7.1.10 虚假数字人脸肤色一致性

应能支持评估虚假数字人脸图像中人脸的额头和左右面颊等子区域之间的肤色的一致性和均衡度。可采用色彩直方图或亮度分布图的相似性来进行该指标的评估。

注：目前的某些虚假数字人脸制作技术的和谐化能力不够会导致面部肤色分布不均的破绽，通过判断人脸不同皮肤子区域之间的肤色一致性，可以得到虚假数字人脸的和谐化程度。虚假数字人脸肤色的一致性越高，相应的攻击能力就会越强。

7.1.11 虚假数字人脸轮廓融合度

应能支持评估虚假数字人脸图像中人脸外轮廓，尤其下颌缘线的清晰度和唯一性。可采用人脸轮廓线的几何特征来进行该指标的评估。

注：制作虚假数字人脸时有面部融合（Blending）的操作，可能在人脸边缘处出现缝合或者伪影的伪造线索。通过人脸边缘轮廓的融合度来重点判断虚假数字人脸的融合操作是否真实自然，融合度越高，伪造痕迹越弱，对应

的攻击能力就会越强。

7.1.12 虚假数字人脸内容前背景色彩分布一致性

应支持评估人脸前景区域和除人脸以外的背景区域之间的色彩、亮度等分布的一致性。可采用色彩直方图或亮度分布图的相似性来进行该指标的评估。

注：制作虚假数字人脸时，无论是基于GAN还是基于Diffusion的生成方法，都可能因为全局和谐化能力不够而导致前后背景在色彩、亮度、光照等方面出现明显差异的问题。当虚假数字人脸内容图像中前背景色彩分布的一致性越高，伪造生成的痕迹就越弱，对应的攻击能力就会越强。

7.1.13 虚假数字人脸内容前背景频谱分布一致性

应能支持评估人脸前景区域和除人脸以外的背景区域之间的频谱特征分布的一致性。可采用傅里叶变换（FFT）后高、中、低频谱段分布图的相似性来进行该指标的评估。

注：制作虚假数字人脸时，无论是基于GAN还是基于Diffusion的生成方法，都可能在篡改区域留下异常的高频噪声线索，导致篡改区域和背景区域的频谱分布出现较大差异化。当虚假数字人脸图像中前背景频谱分布的一致性越高，伪造生成的痕迹就越弱，对应的攻击能力就会越强。

7.1.14 虚假数字人脸内容目标身份 ID 保真度

应能支持评估虚假数字人脸图像中的人脸ID和目标人脸ID之间身份信息的相似度。可使用人脸比对模型对换脸后的人脸和目标人脸之间的身份特征进行比对，比对分越高就表示目标身份ID的保真度越好。

注：制作虚假数字人脸时，攻击方通常期望生成时的身份ID高度可控、并将目标身份ID没有丢失的完全替换到图像中，既能增加篡改后身份信息的逼真度，也能提高对人脸识别系统全链路的攻击能力。当虚假数字人脸图像中目标身份ID的保真度越高，对应的生成虚假数字人脸就越真实，其攻击能力就会越强。

7.1.15 基于人脸活化类技术制作的虚假数字人脸视频的驱动源位姿保持度

应能支持评估虚假数字人脸视频中每帧图像的人脸位姿信息和驱动源视频中每个人脸动作的位姿信息之间的一致性。可使用虚假数字人脸和驱动源人脸之间的头部空间位置差异值HPMSE或者旋转角度差异值PRMSE来进行该指标的衡量。

注：制作虚假数字人脸视频目标是能够将驱动源动作的位姿更加精准无损地迁移到被驱动的人脸图像上，但有时会因为可控力度和生成精度不够而出现位姿保持不准、动作变形异常等问题。当虚假数字人脸视频中每帧图像的人脸位姿和驱动源位姿的保持度越高，说明位姿模拟更加精准、运动信息更加仿真，相应的攻击能力就会越强。

7.1.16 虚假数字人脸视频的单帧人脸几何协调度

应能支持评估虚假数字人脸视频中每帧图像内部虚假数字人脸的五官位置之间的几何协调度，可采用虚假数字人脸关键点和单元模板脸关键点之间的平均关键点距离值AKD来进行该指标的衡量。

注：制作虚假数字人脸视频时，有概率会出现重建精度不够而出现面部五官变形、表情扭曲和轮廓线条不流畅的问题，从而暴露伪造线索。当虚假数字人脸视频中每帧图像内部的虚假数字人脸在五官分布上的几何协调度越高，说明虚假数字人脸视频更加自然流畅，相应的攻击能力就会越强。

7.1.17 虚假数字人脸视频的帧间人脸身份 ID 一致性

应能支持评估虚假数字人脸视频中不同图像帧之间的同一虚假数字人脸在身份ID特征上的一致性。

注：制作虚假数字人脸视频过程中，在驱动人脸执行不同动作时可能在时间维度出现人脸ID特征漂移和失真等不稳

定的问题，从而被防御模型识破。当虚假数字人脸视频中不同图像帧之间的人脸身份一致性越高，说明虚假数字人脸的身份信息更加真实鲁棒，相应的攻击能力就会越强。

可引入基于深度学习的人脸比对模型对不同帧之间的人脸身份特征进行比对，比对分越高就表示不同帧之间的人脸身份一致性越好。

7.1.18 虚假数字人脸视频的帧间人脸动作连续性

应能支持评估虚假数字人脸视频中不同图像帧之间的同一虚假数字人脸在动作位姿和运动信息上的连续性。可采用帧间人脸的位置、角度、速度等运动量的变化速度来进行评估，也可采用Farneback稠密光流场进行更精细化的运动连续性建模。

注：制作虚假数字人脸视频过程会由驱动源去执行不同难度的动作，往往会因为动作幅度大小和人脸重建精度而出现一些动作变形僵硬、缺乏深度立体感，或者运动异常导致的视频断层和抖动问题，这些都是容易被检出的虚假线索。当虚假数字人脸视频中不同图像帧之间的人脸动作连续性越高，说明虚假数字人脸的运动信息更加流畅合理，相应的攻击能力就会越强。

7.1.19 虚假数字人脸视频的帧间人脸光线差异性

应能支持评估虚假数字人脸视频中不同图像帧之间的同一虚假数字人脸在光线、亮度上的差异性，可采用帧间人脸光线直方图或亮度分布图的相似性来进行该指标的评估。

注：制作虚假数字人脸视频的人脸由于同源像素拷贝性太高而会出现视频中人脸表情呆滞、面部光线阴影一成不变的问题，这和真实人脸具备正常光影变化的情况之间有着较大差异。当虚假数字人脸视频中不同图像帧之间的人脸光线分布存在正常阈值范围内的差异性时，说明虚假数字人脸具备更自然的光影变化和微表情，相应的攻击能力就会越强。

7.1.20 使用手工篡改类技术制作的虚假数字人脸内容的篡改区域融合度

应能支持评估指虚假数字人脸内容篡改区域和未篡改区域之间在肤色、亮度和边缘等方面的融合度。可采用篡改区域和未篡改区域在色彩直方图或亮度分布图上的相似性来进行该指标的评估。使用人脸属性编辑类技术制作的虚假数字人脸内容的可编辑属性类型。

注：手工篡改应进行区域融合（Blending）的后处理操作，可能在篡改边缘处出现缝合或者伪影的伪造线索。人脸篡改区域和未篡改区域之间的融合度越高，融合操作更加真实和谐，伪造痕迹越弱，对应的攻击能力就会越强。

7.2 性能要求

7.2.1 概述

虚假数字人脸内容检测服务性能可从准确性、鲁棒性、泛化性、响应速度等四个维度进行评估，具体评估指标及查验内容见表4。

表 4 性能要求评估指标及查验内容

指标类别	指标项	查验内容	必选/可选
性能要求	准确性	考察虚假数字人脸内容检测服务面向不同类型深度伪造和数字化攻击	必选

		的检测准确程度	
	鲁棒性	考察虚假数字人脸内容检测服务面向虚假数字人脸内容叠加不同干扰的鲁棒能力	可选
	泛化性	考察虚假数字人脸内容检测服务面向虚假数字人脸内容扩展新方法、新数据的泛化能力	可选
	响应速度	考察虚假数字人脸内容检测服务处理图像的吞吐量	必选
		考察虚假数字人脸内容检测服务处理每秒视频单元的吞吐量	必选

7.2.2 准确性

准确性通过真假人脸检测的准确率和虚假人脸的召回率双指标来衡量。

a) 准确率的定义是指正确检测的样本数占总样本数的比例，该指标值越高表示虚假数字人脸内容检测服务对真假判断的准确性越好，指标计算公式如（1）所示：

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(1)}$$

式中：

TP——检测服务将虚假人脸正确识别为假的样本数；

TN——检测服务将真实人脸正确识别为真的样本数；

FP——检测服务将真实人脸错误识别为假的样本数；

FN——检测服务将虚假人脸错误识别为真的样本数；

b) 召回率是指固定真实人脸通过率下对虚假人脸的召回率，该指标值越高表示虚假数字人脸内容检测服务在合理真通过率下对防御虚假人脸的准确性越好，指标计算公式如（2）所示：

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(2)}$$

式中：

TP——检测服务将虚假人脸正确识别为假的样本数；

FN——检测服务将虚假人脸错误识别为真的样本数；

固定合理真通过率下的假召回率指标计算流程如下：

a) 检测服务输出对所有待测样本预测的风险概率分，并分别对真假样本的风险概率分进行从高到低排序；

b) 给定合理真通过率 P（推荐值为 90%），取 P 所对应的风险概率分作为拦截阈值 Thresh，该阈值用来保证真实人脸的通过率为 P。真通过率即真实人脸样本被正确预测为真的数量占真实人脸样本总数量的比例。

c) 将阈值 Thresh 应用到所有的假样本上，即可计算出假召回率 R。假召回率即虚假人脸样本被

正确预测为假的数量占虚假人脸样本总数量的比例。

d) 此时计算出的假召回率 R 即为固定真通过率下的假召回率，准确性指标由 R 来衡量。

7.2.3 鲁棒性

鲁棒性通过更换叠加干扰数据后准确性指标的退化率来衡量，退化率的定义是更换叠加干扰数据前后准确性指标的偏差绝对值与更换数据前的准确性基准值之比，退化率越小说明虚假数字人脸内容检测服务对抗干扰的鲁棒性越好，其计算公式如（3）所示：

$$D_r = \frac{|Acc - Acc\_base|}{Acc\_base} \quad \text{.....(3)}$$

式中：

D_r——更换叠加干扰数据后检测服务的准确性指标的退化率；

Acc——更换叠加干扰数据后检测服务的准确率；

Acc_base——更换叠加干扰数据前检测服务的准确率基准值。

7.2.4 泛化性

泛化性通过更换新攻击数据后准确性指标的退化率来衡量，退化率的定义是更换新攻击数据前后准确性指标的偏差绝对值与更换数据前的准确性基准值之比，退化率越小说明虚假数字人脸内容检测服务应对新型攻击的泛化性越好，其计算公式如（4）所示：

$$\mathrm{D_g} = \frac{|\mathrm{Acc}^{'} - \mathrm{Acc\_base}^{'}|}{\mathrm{Acc\_base}^{'}} \quad \text{.....(4)}$$

式中：

D_g——更换新攻击数据后检测服务的准确性指标的退化率；

$Acc^{'}$——更换新攻击数据后检测服务的准确率；

$Acc\_base^{'}$——更换新攻击数据前检测服务的准确率基准值。

7.2.5 响应速度

响应速度通过服务系统处理图像的吞吐量和处理每秒视频单元的吞吐量双指标来衡量。

a) 处理图像的吞吐量，其定义是检测服务平均每秒所能处理的人脸图像样本总量，计算公式如（5）所示：

$$QPS_{image} = \frac{N}{T_e - T_s} \quad \text{.....(5)}$$

式中：

T_s——参评服务开始运行时刻；

T_e——图像样本全部检测完成后，参评服务停止运行时刻；

N——数据集图像样本总量；

QPS_{image}——参评服务处理图像的吞吐量。

b) 处理每秒视频单元的吞吐量，其定义是检测服务平均每秒所能处理的每秒视频单元总量，计算公式如（6）所示：

$$QPS_{video} = \frac{\sum_{i=1}^{N} T_i}{T_e - T_s} \quad \text{.....(6)}$$

式中：

T_s——参评服务开始运行时刻；

T_e——视频样本全部检测完成后，参评服务停止运行时刻；

N——数据集视频样本总量；

T_i——第i个视频样本时长，以秒为单位；

QPS_{video}——参评服务处理每秒视频单元的吞吐量。

7.2.6 评估方法

虚假数字人脸内容检测服务的性能评估主要依托服务提供的数据接口并且基于评测数据集展开测试，按照各指标计算方法进行统计和分析。各性能指标对应评测数据说明参见7.1。

8 能力分级和评估方法

金融虚假数字人脸内容检测服务评估等级划分为依次递增的五个能力等级：基础级（一级）、增强级（二级）、优秀级（三级）、卓越级（四级）和引领级（五级）。对应能力等级的具体要求描述如下：

a) 基础级（一级）：虚假数字人脸内容检测服务应具备基础的检测功能和服务能力。基本满足功能和服务能力的各项必选指标；

b) 增强级（二级）：在基础级的基础上，虚假数字人脸内容检测满足增加的能力指标；

c) 优秀级（三级）：在增强级的基础上，虚假数字人脸内容检测应具备性能要求，部分指标应达到行业领先水平；

d) 卓越级（四级）：在优秀级的基础上，虚假数字人脸内容检测应具备更高的性能，各项指标基本达到行业领先水平；

e) 引领级（五级）：在卓越级的基础上，虚假数字人脸内容检测应满足所有能力项。性能要求的所有指标应达到行业领先水平。

参 考 文 献

[1] GB/T 25069—2022 信息安全技术 术语
[2] GB/T 37036.8-2022 信息技术 移动设备生物特征识别 第8部分：呈现攻击检测
[3] GB/T 41815.1-2022 信息技术 生物特征识别呈现攻击检测 第1部分：框架
[4] GB/T 41987-2022 公共安全 人脸识别应用 防假体呈现攻击测试方法