ICS 35.040
L 71

团 体 标 准

T/AI 110.2—2022

人工智能视觉隐私保护
第 2 部分：技术应用指南

Visual privacy protection of artificial intelligence—
Part 2:Technique application guide

2022－12－30 发布　　2022－12－30 实施

中 关 村 视 听 产 业 技 术 创 新 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。

本文件是T/AI 110《人工智能视觉隐私保护》的第2部分。T/AI 110已经发布了以下部分：

——第1部分：通用技术要求

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专利的责任。

本文件由新一代人工智能产业技术创新战略联盟AI标准工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：海信集团控股股份有限公司、山东大学、深圳龙岗智能视听研究院、美的集团(上海)有限公司、中国海洋大学、天津大学、上海商汤智能科技有限公司、云从科技集团股份有限公司、翼健（上海）信息科技有限公司、上海数字电视国家工程研究中心有限公司、北京邮电大学、西安电子科技大学、北京大学、北京聪明核桃教育科技有限公司。

本文件主要起草人：陈维强、高雪松、刘琚、张世雄、区志财、魏志强、李克秋、张淯易、李亚锋、魏文应、唐剑、刘秀龙、徐浩、李军、周禾、顾凌晨、刘晓玺、王雪静、邵宸、刘常昱、吕林、黎俊良、殷惠清、陆月明、朱辉、田永鸿、王荣刚、牟小峰、蒋慧、吴庚、温浩、尹建华、左金鑫、蔡亚森、鲁昱、赵雪圻、陶键源、李若尘、李哲。

引　言

隐私数据是指数据中直接或间接蕴含的，涉及个人或组织的，不宜公开的，需要在收集、存储、查询和分析、发布过程中加以保护的信息，而视觉隐私数据在隐私数据的基础上具有视觉特征，包括视频、图像等。T/AI 110旨在确立适用于人工智能视觉隐私保护的设计、应用、评估等方面遵循的原则和相关规则，拟由3个部分构成。

——第1部分：通用技术要求。目的在于确立人工智能视觉隐私保护设计、实施等所需要遵守的总体要求。

——第2部分：技术应用指南。目的在于对人工智能视觉隐私保护技术的应用提供指导。

——第3部分：测试评估。目的确立适用于人工智能隐私保护的测试评估。

本文件对保护视觉隐私数据的算法技术提出基本建议，它们符合用户对视觉隐私数据广义上的认知，同时也对隐私数据算法的性能提出基本建议。本文件只涉及视觉隐私数据保护的技术与处理，现阶段仅暂用于家庭监控视频中。

本文件的目的是对视觉隐私数据保护算法做出最适合的建议和最佳的科学操作指导。目前，针对视觉数据保护算法的度量缺乏统一的标准及行业性建议。现存的安全性要求都是从算法本身的衡量指标角度出发，过于笼统。视觉隐私数据保护算法的种类有很多，每一个算法重点保护的对象不同，不同算法产生的代价与带来的保护收益不同导致准确度与性能差异很大。这些差异导致本文件无法产生准确统一的安全性与可用性约束条件。因此，本文件抽象出视觉隐私保护算法的共性，针对算法的性能以及代价不同进行了统一的定性，其理念和原则广泛适用于视觉隐私数据保护算法。

人工智能视觉隐私保护 第2部分：技术应用指南

1 范围

本文件给出了视觉数据的保护目标和措施，提供了人工智能视觉隐私保护技术应用过程中的管理措施指南。

本文件适用于企业对海量视觉数据隐私保护措施的设计、应用等，也适用于网络安全相关主管部门、第三方评估机构等组织开展视觉隐私数据的安全监督管理、评估。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069—2022 信息安全技术 术语

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 37964—2019 信息安全技术 个人信息去标识化指南

T/AI 110.1—2020 人工智能视觉隐私保护 第1部分：通用技术要求

3 术语和定义、缩略语

3.1 术语和定义

GB/T 25069—2022和GB/T 35273—2020界定的以及下列术语和定义适用于本文件。

3.1.1

视觉隐私数据 visual privacy data

视觉数据中直接或间接蕴含的，涉及个人或组织的，不宜公开的，需要在数据收集、数据存储、数据查询和分析、数据发布等过程中加以保护的信息。

3.1.2

实例分割 instance segmentation

对视觉数据进行有效的分割，并进行保护的算法。

3.1.3

差分隐私 differential privacy

从统计数据库查询时，最大化数据查询的准确性，同时最大限度减少识别其记录的机会。

3.1.4

结构化数据 structured data

由二维表结构来逻辑表达和实现的数据，严格地遵循数据格式与长度规范，主要通过关系型数据库进行存储和管理。

注：针对原始数据进行过预处理的数据，可结构化为哈希码、权重矩阵等形式。

3.1.5

非结构化数据 unstructured data

不适于由数据库二维表来表现的数据。

注：针对原始的、未经过任何处理即可直接使用的数据，与3.1.4结构化数据相对应。

3.1.6

信息隐藏 information hiding

对视觉数据进行特定信息的嵌入，实现数据隐私保护的算法。

3.1.7

小样本检测 few-shot detection

融合传统的目标检测与小样本学习方法，基于包含充足标注数据的基类，通过训练方法设计、模型结构设计与损失函数设计，引导模型在极少量标注数据中学习具有一定泛化性能的检测模型。

3.2 缩略语

下列缩略语适用于本文件。

AP：平均精度(Average Precision)
BPP：每像素位数(Bits Per Pixel)
FPS：每秒传输帧数(Frames Per Second)
GAN：生成对抗网络(Generative Adversarial Network)
KID：内核感知距离(Kernel Inception Distance)
mAP：平均精度均值(mean Average Precision)
MMD：最大均值误差(Maximum Mean Discrepancy)
SSIM：结构相似性(Structural Similarity Index)
PSNR：峰值信噪比(Peak Signal-to-Noise Ratio)

4 视觉隐私保护算法技术总则

4.1 隐私数据控制原则

对于人工智能视觉隐私数据的控制者，其所应履行的义务，采取的安全保障措施等参照T/AI 110.1—2020中第4章。

4.2 算法通用性原则

4.2.1 脱敏性

根据使用方的需求，对原始的视觉数据实现不同的目标，扰动、模糊或替换视频与图像中的人脸，身份标识、个人物品与行为等可识别个人属性的信息，有效的保护个体隐私内容从而实现脱敏，防止原始数据被恶意窃取与非法利用。

4.2.2 可利用性

针对脱敏后的隐私数据，在最大限度地保护隐私数据的前提下，保证不影响处理后数据的正常使用。本文件建议，用户数据信息的减少宜在限定的可控范围内，且不影响使用方的正常使用；计算复杂度的增加同样宜控制在使用方可接受的范围内。除此以外，存储开销宜控制在服务操作可用的范围内，使用者可更高效地使用经处理的视觉数据。

4.2.3 鲁棒性

不同的数据隐私保护模型在出现异常时保持其算法性能的能力，一般包括环境变化或系统安全受到威胁等异常情况。当模型假设在可接受的范围内出现偏差，对算法性能产生的影响仍能满足正常使用的需求。降低由于异常情况造成的算法性能下降，保证算法能够有效的进行隐私数据保护。

4.3 算法用户友好性原则

4.3.1 可恢复性

隐私保护算法执行前后，隐私信息被还原的能力。安全算法的不可逆性是十分必要的，即脱敏处理后的数据无法推断出用户原始的数据信息。考虑用户实际应用需求，对视频授权者可保留算法可恢复性，即通过逆向算法获取带有原始隐私信息的视觉数据。对于个人信息的处理参照GB/T 37964—2019的规定。

4.3.2 实时性

为满足用户对视觉数据的使用需求，对于视频数据的可用算法推理速度不应低于25 FPS，避免视频出现卡顿，保证视频数据的显示效果。

4.3.3 泛化性

隐私数据保护模型应考虑隐私个体主观性与场景差异性，根据用户自身需求，满足对不同类别的隐私内容的检测与保护处理，保证模型泛化能力与适应性。

5 视觉隐私保护算法评价指标

5.1 平均精度均值

不同类别精确率的平均值，为实例分割、小样本检测、图像缩放等技术的常用指标。如用来衡量对于某一个实例定位是否准确，以此来判断对隐私数据的脱敏性保护效果。

AP用于衡量某一类别的误差，以查全率为横轴，查准率为纵轴，对构成的曲线进行积分计算。计算公式如式(1)所示：

$$AP = \int_0^1 P(R)dR \quad \cdots\cdots (1)$$

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

式中：

P——查准率；

R——查全率；

TP——正确预测出的正样本数量；

FP——负样本被预测为正样本的数量；

FN——正样本被预测为负样本的数量。

参照现有算法的实现效果，mAP应达到25%以上，计算公式如式(2)所示。

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP \quad \cdots\cdots (2)$$

式中：

Q——不同类别的数量；

AP——平均精度。

5.2 结构相似性

图像数据脱敏可利用性的衡量指标。通过图像的亮度、对比度和结构，衡量两幅图像的相似性。用平均混度作亮度测量，灰度标准差则作对比度测量。图像结构差异计算公式见式(3)。

$$s(x,y) = \frac{\sigma_{xy} + C_1}{\sigma_x \sigma_y + C_1} \quad \cdots\cdots (3)$$

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2}$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$$

式中：

$s(x,y)$——两幅图像的结构差异；

σ_x——图像 x 的标准差；

σ_y——图像 y 的标准差；

σ_{xy}——图像 x 和 y 的协方差；

C_1——常数；

N——图像像素点的数量；

μ_x——图像 x 的均值；

μ_y——图像 y 的均值。

SSIM的标准计算公式见式(4)。为满足图像可利用性，SSIM不应低于0.50，确保肉眼观察无明显差别。

$$SSIM(x,y)=[l(x,y)^{\alpha}c(x,y)^{\beta}s(x,y)^{\gamma}] \quad \cdots\cdots (4)$$

$$l(x,y)=\frac{2\mu_x\mu_y+C_2}{\mu_x^{\ 2}+\mu_y^{\ 2}+C_2}$$

$$c(x,y)=\frac{2\sigma_x\sigma_y+C_3}{\sigma_x^{\ 2}+\sigma_y^{\ 2}+C_3}$$

式中：

$l(x,y)$——两幅图像的亮度差异；

$c(x,y)$——两幅图像的对比度差异；

$s(x,y)$——两幅图像的结构差异，计算公式如式(3)所示；

α 、β 和 γ——常数；

C_2 和 C_3——常数。

5.3 内核感知距离

图像数据脱敏可利用性的衡量指标。KID是在Inception网络上真实图像和生成样本之间的MMD，是一种无偏估计器，对于少量测试样本也较有效。该距离越小，代表生成效果越好。当KID分数小于8时，即可认定生成较好，满足可利用性。

5.4 峰值信噪比

嵌入数字水印后图像质量的测量指标。通过信号最大有效功率和噪声功率的比值，衡量生成图像的质量。PSNR常用对数分贝为单位，根据现有算法性能，峰值信噪比不应低于30 dB，否则图像不具备可利用性。

5.5 信息隐藏容量

图像内嵌入数字水印的信息大小。信息隐藏容量用于衡量原始图像或视频内嵌入信息的大小。单位为BPP，在保证图像不失真的情况下，数值越大，代表载体信息容量越高。

5.6 每秒传输帧数

实时性衡量指标。FPS用于测量动态视频的信息数量。每秒传输帧数越多，所显示的画面越流畅。在家庭环境下，摄像头典型帧率为25 FPS，低于25 FPS会影响录像视频的流畅度。为不影响基于原始视觉数据的隐私保护使用，可用算法推理测试不应低于25 FPS。

5.7 元任务类别与样本数

衡量隐私保护模型泛化性的指标。小样本数据含有多个类别且每个类别包含有限个标注样本，构成一个模型一个元任务。通过不同的元任务组合来保证模型泛化能力与灵活性。通常，在标注数据数量越少的情况下，可识别的样本类别越多，代表模型的泛化性越强。

6 基于非结构化视觉数据的隐私保护要求

6.1 基于风格迁移的隐私保护技术

风格迁移也称为图像风格化，目的是保证图像在内容不变的情况下实现图像的风格转化，技术使用要求如下：

a) 在视觉效果上，生成类似原始图像风格的伪图，为保证图像可利用性，SSIM 不应低于 0.50，KID 不应高于 8；
b) 根据用户实时性需求，技术处理不应低于 25 FPS，以保证视频隐私内容实时性；
c) 应满足替换目标的可选择性与处理后内容的可利用性；
d) 根据用户需求选择是否对授权者保留算法可逆性，通过特定方式获取最原始视频信息。如原始视频中包含个人信息，对该视频的处理参照 GB/T 35273—2020 的规定；
e) 在指定对象前提下，有针对性的替换保护，在公共信道中传输以达到混淆第三方视听的效果；
f) 本方法使用前，需使用目标检测方法对图像或视频帧中的待处理区域进行提取。

注：常用的效果较好的目标检测方法有YOLO、RetinaNet、SSD等。

6.2 基于实例分割的隐私保护技术

实例分割能够区分不同对象并提取轮廓信息，实现了针对不同对象的保护，技术使用要求如下：

a) 在视觉效果上，通过对隐私信息模糊或遮挡处理实现隐私内容的脱敏；
b) 根据用户实时性需求，技术处理不应低于 25 FPS，以满足视频隐私内容实时性；
c) 该技术应达到业务所需平均精度均值，以保证隐私内容的准确定位；
d) 隐私信息的遮挡或模糊处理，保证原始数据的整体可利用性。
e) 具备一定的抵御对抗样本攻击能力，以提高训练模型的鲁棒性；
f) 在亮度、对比度、饱和度发生轻微变化，或有轻微噪声输入的情况下，功能仍可正常使用。

6.3 基于信息隐藏的隐私保护技术

信息隐藏将隐私数据隐藏于可公开的媒体信息中，具有人眼无法察觉的优点，能够有效对隐私数据的保护，技术使用要求如下：

a) 在视觉效果上，该技术生成的图像与原始图像相似，SSIM 不应低于 0.50，PSNR 不应低于 30 dB；
b) 根据用户实时性需求，技术处理不应低于25 FPS，以满足视频隐私内容实时性；
c) 应保证生成图像内含信息满足不可视性；
d) 应能够完整的提取出隐私数据，以保证隐私数据的可利用性和算法的鲁棒性；
e) 宜可嵌入尽可能多的隐私数据，满足信息容量的需求。
f) 算法具备一定的鲁棒性，图像在正常的滤波、编解码、数模转换后，仍可完整的提取出隐私数据。

6.4 基于小样本检测的隐私保护技术

针对数据处理成本高的隐私内容，融合传统目标检测与小样本学习技术，引导模型在极少量数据集基础上学习得到具有一定泛化能力的隐私内容检测模型，从而实现对其相应的保护处理，技术使用要求如下：

a) 在视觉效果上，实现对隐私信息模糊或遮挡(替换)处理，满足隐私内容的脱敏；
b) 根据用户实时性需求，技术处理不应低于 25 FPS，以满足视频隐私内容实时性；
c) 模型对于新的外部数据宜保持性能稳定，以满足模型的鲁棒性；
d) 应达到业务所需平均精度均值，以保证隐私内容的准确定位；
e) 在元任务组合中宜满足用户实际应用需求，保证对小样本隐私内容的检测准确性与模型泛化性；
f) 模型宜考虑隐私内容的个体主观性与场景的差异性，以满足模型的灵活应用；
g) 隐私信息的遮挡或模糊处理，宜满足原始数据的整体可利用性。

6.5 基于图像缩放的隐私保护技术

图像缩放技术，通过高倍数的图像降采样，有效对隐私数据进行保护；并可在需要时使用超分辨技术，还原数据的绝大部分特征和细节，实现部分数据可逆恢复。技术使用要求如下：

a) 在视觉效果上，通过对隐私信息模糊或遮挡处理实现隐私内容的脱敏；
b) 根据用户实时性需求，技术处理不应低于 25 FPS，以满足视频隐私内容实时性；
c) 该技术应达到平均精度均值，以保证隐私内容的准确定位；
d) 隐私信息的遮挡或模糊处理，保证原始数据的整体可利用性；
e) 模型宜考虑隐私内容的使用场景差异，以满足模型的灵活应用。

7 基于结构化视觉数据要求

7.1 原始数据的降维处理及显著性表征

针对存储空间的限制和个人隐私问题，典型的方法是对原始数据做二值化处理操作，并采用深度神经网络构建映射关系：

a) 数据通过基于深度学习的哈希方法，构建映射关系；
b) 数据存在与其唯一对应的标签；
c) 数据通过二值码来表示；
d) 数据的相似度度量为数据二值码之间的汉明距离；
e) 数据二值码宜具备高匹配准确度特性。

7.2 差分隐私和特征数据保护

通过对数据添加噪声，有效抵抗差分攻击，降低了隐私数据的敏感程度：

a) 该技术针对不同的差分攻击具备相应的抵御能力；
b) 在数据保护处理过程中，噪声添加等特殊方式达到隐私数据低敏感度；
c) 差分隐私方法可以在原始数据二值化处理的基础上实施。

7.3 基于通用事件流的隐私保护方法

通用事件流只记录原始的视频数据中运动目标的轮廓信息，忽略了视频处理中的关键因素时序特征的描述，能够保护视频中的隐私数据：

a) 在视觉效果上，宜实现对事件数据充分稀疏化处理的同时凸显运动特性，以满足图像可利用性；
b) 根据用户实时性需求，技术处理不应低于 25 FPS，以满足视频隐私内容实时；
c) 该技术应保证生成图像、视频文件保留足够信息可供人眼查验和数据标注；
d) 该技术宜对视频数据进行稀疏压缩处理，以便事件传输、存储和分析；
e) 该技术提取的特征宜支持多种颜色系统编码。

附 录 A
（资料性）
隐私保护方法简介

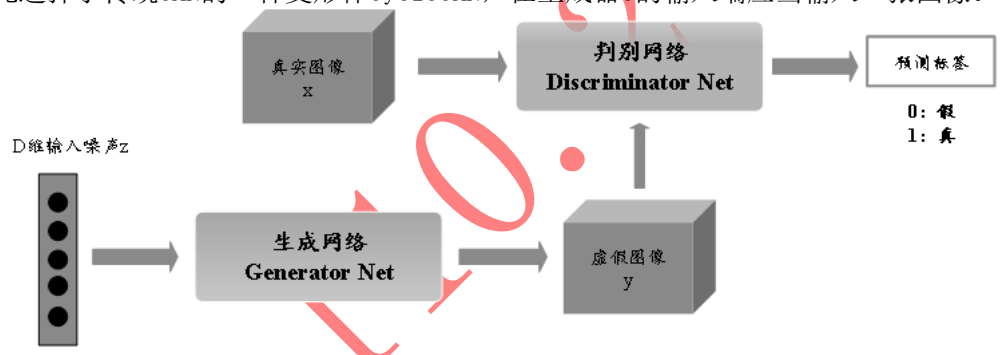
A.1 基于非结构化视觉数据的隐私保护方法

A.1.1 总则

针对于原始的视频数据，本文件提出替换与模糊两类方法以满足标准，其他能够满足该文件原则的方法同样可以纳入其中。

A.1.2 基于风格迁移的隐私保护算法

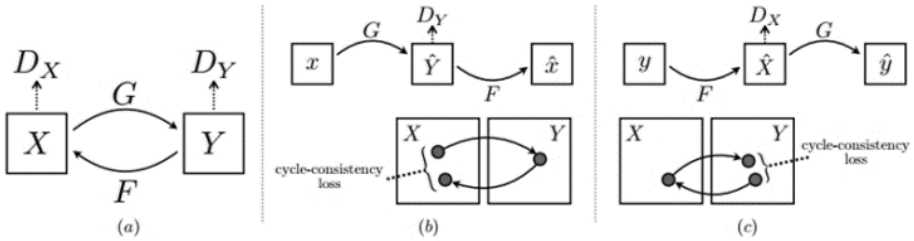
替换原始数据信息方法，主要针对人脸信息保护。目前主流的生成模型有流模型、自回归模型和生成对抗网络(GAN)三种。生成对抗网络是通过生成器和判别器两个独立的神经网络，对输入的任意维度噪声生成视觉合理的图像，从而达到混淆视听的目的。考虑到生成的高效性和并行性，本文件在针对原始非结构化数据的伪图生成中使用了GAN网络，且GAN网络是不可逆的。传统GAN是通过生成网络G(Generator)和判别网络D(Discriminator)不断博弈，进而使G学习到真实数据的分布。将其应用到图片生成上，则训练完成后，G可以从一段随机噪声中生成逼真的图像，见图A.1。但是，考虑到数据隐私中涉及到的都是图像和视频等数据，本文件期望对原始视觉隐私数据进行脱敏或加密后达到混淆视听的效果。因此选择了传统GAN的一种变形体CycleGAN，在生成器G的输入端应当输入一张图像。



图A.1 GAN 示意图

该网络和传统GAN的不同在于，G接收的不再是一个随机噪声x，而是一张视频监控中的原始图像，G(x)为生成图像，输出的D(x)代表生成图像为真实图像的概率。训练中，生成网络G的目标就是尽量生成真实的图片去欺骗判别网络D，而D的目标就是努力判别输入网络的是真实图像还是生成伪图。这样G和D就构成了一个动态的“博弈过程”，最终的平衡点即纳什衡点。

CycleGAN网络在本质是两个镜像对称的单向GAN网络。两个单向的GAN共享两个生成器，并各自带一个判别器，即共有两个判别器和两个生成器。输入的非成对图像x和y，通过训练创建映射，确保输入输出可以共享一些特征，见图A.2。在视觉隐私数据保护任务中，可以生成类似原始图像风格的伪图，实现保护当事人隐私的目标。



图A.2 CycleGAN 示意图

与CycleGAN同期的两个模型DualGAN和DiscoGAN，实际上都采用了和CycleGAN一样的思路，即利用循环生成的思想，输入样本连续经过两个对称的生成网络后，期望输出一个和输入完全一致的图像，因此在损失函数方面也采用了循环一致的L1范数。但是，由于面向的具体任务不同，三者在网络结构上存在细微差别。DualGAN采用了U-net的网络结构，同时也包含和CycleGAN中相同的残差块。而DiscoGAN结构更近似于DCGAN。CycleGAN更强调非成对图像间的翻译任务，且多个评价指标的衡量中，CycleGAN均在三者中具有最好性能。表A.1展示了三种模型在FID(Fréchet Inception Distance)、KID以及SSIM分数上的对比。

表A.1 三种模型的量化对比

评价指标 / 模型	FID	KID	SSIM
DualGAN	140	10	0.41
DiscoGAN	145	11	0.37
CycleGAN	120	7	0.50

在实时性需求上，常见的视频的帧率在25-30 FPS之间，通常GAN的图像生成速度在20-30 ms之间，因此可以满足视频隐私实时性与脱敏性。

同时，与现有的一些通过对图像视频等数据中隐私部位直接打马赛克的方式相比，本文件的方法具有以下优势：如果需要替换，可以指定被替换对象和待替换目标，通过定向替换，一是可以在传递信息过程中，只可以看到被替换后的目标，达到混淆视听的目的；二是有权限的视频处理人员和被授权者可以通过替换关系，获取最原始的视频信息。

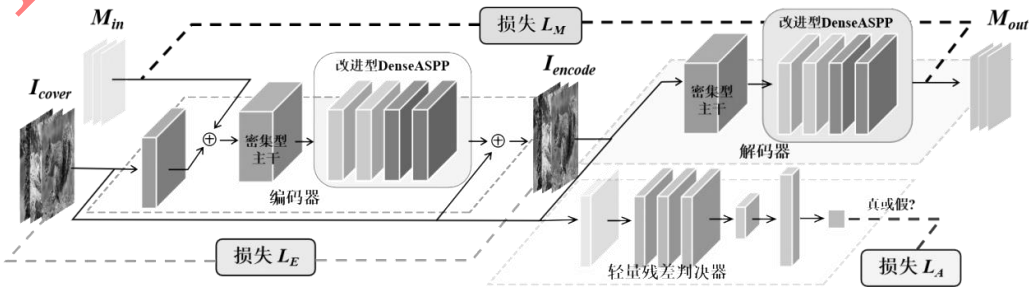
需要注意的是，本方法在使用前，需使用目标检测方法对图像或视频帧中的待处理区域进行提取。常用的效果较好的目标检测方法有YOLO、RetinaNet、SSD等，本文件不做具体要求。

A.1.3 基于实例分割的隐私保护技术

模糊原始数据信息方法，主要针对实物隐私保护。使用编码器与解码器两类方法进行搭建，编码器使用深度卷积神经网络进行提取特征，其作用为能够实现在不同图像尺度下进行特征信息的提取与融合，例如使用深度可分离卷积等，而解码器的作用为将提取的高维特征进行分类恢复为原始图像尺寸的大小，例如使用U-net网络等。使用图像分割的思想能够选择性的保护需要保护的区域，而模糊掉不需要的或者敏感的信息。该算法能够在较高分辨率的情况下有效地提取出所需要的信息。

A.1.4 基于信息隐藏的隐私保护技术

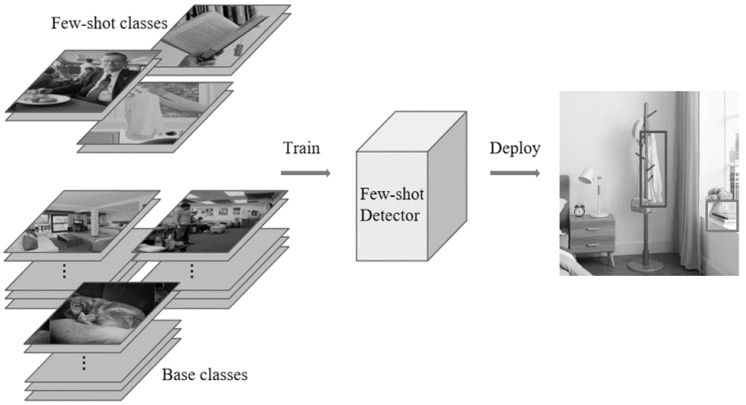
嵌入隐秘信息至原始数据信息，主要针对图像进行安全和隐私保护。如图A.3所示，基于生成对抗式网络，在生成器的编码网络中加入空洞空间金字塔，提取多尺度和多空洞率的特征，大幅度提升感受野范围，从而获取图像的结构冗余信息以提高图像信息隐藏容量和信息的不可视性；为提高上述嵌入信息的提取准确度，保证多方隐秘通信和隐私信息取证鉴定的顺利完成，在生成器的解码网络使用空洞空间金字塔，获取隐藏信息的语义特征和细节特征以解码出所嵌入信息，同时使用二值交叉熵损失来提高解码准确度；在网络训练的过程中使用不同尺度梯度更新策略，加快网络的对抗博弈并且稳定快速达到最佳平衡点，使用余弦退火调整学习率以提高梯度的收敛速度和效率，使用密集型连接方式缓解梯度消失的问题，大幅度减少参数。该算法能够在较高分辨率的情况下，进行信息的嵌入和有效提取，实现多媒体图像的隐私安全保护。



图A.3 算法结构框图

A.1.5 基于小样本检测的隐私保护技术

考虑到隐私保护技术的实际应用需求，主要解决实际中数据收集与标注困难的问题。融合传统的目标检测与小样本学习方法，基于包含充足标注数据的基类，通过训练方法设计、模型结构设计与损失函数设计，引导模型在极少量标注数据中学习具有一定泛化性能的检测模型。保留基类权重模型，对于不同元任务的组合进行参数微调，实现对小样本隐私内容的检测定位与保护处理，进一步提高模型的泛化能力与灵活性，见图A.4。



图A.4 小样本检测示意图

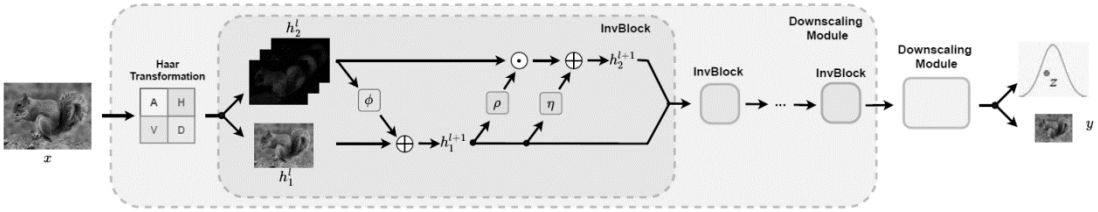
A.1.6 基于图像缩放的隐私保护技术

使用图像缩放算法，能模糊原始数据信息，对实物隐私保护。如图A.5所示，图片经过Haar小波变换后(分离高频和低频)，进入可逆downscale模块，会输出两个结果，一个是缩小后的图片，另一个是丢失的高频成分建模形成的隐分布(与输入图像独立，即与输入图像无关)。当需要恢复原图时，利用低分辨率图像和上述的隐分布数据就可以重建原来高分辨的图像。使用者可以根据应用场景，选择不同倍数的图像缩小。本方法可以直接应用在全画面，加密时对全图进行降采样，还原时采用可逆图像缩放算法。

在实时性需求上，图像可逆缩放算法的推理速度经过优化后，可达100-200 ms，可满足低码率视频隐私实时性与脱敏性。

在实践中，也可先使用目标检测方法对需要加密的区域提取，再应用(可逆)图像缩放技术，这样可以进一步提高处理速度，有利处理高分辨率数据。常用的目标检测有yolo、SSD、RetinaNet、EfficientDet等，本文件不做具体要求。

也可以考虑多级缩放的技术方案，即每次缩放为X4或X2倍，这样的话则需要采用多个(可逆)图像缩放模型串联。



图A.5 可逆图像缩放算法示意图

A.2 基于结构化视觉数据技术方法

A.2.1 总则

针对文本等结构化数据，本文件提出原始数据预处理及表征方法，在此基础上实现结构化数据的隐私保护，其他能够满足该文件原则的方法同样可以纳入其中。

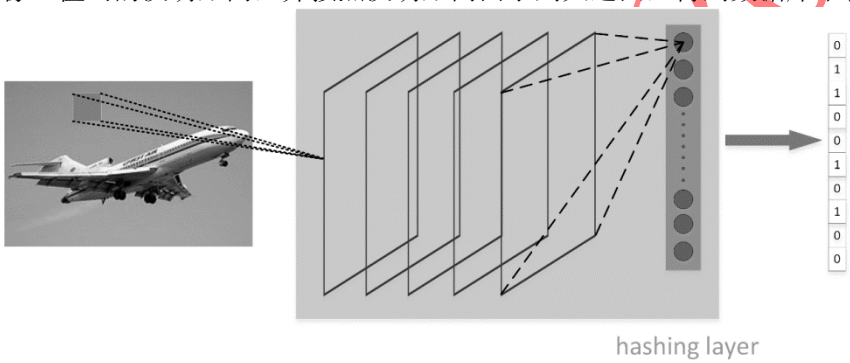
A.2.2 原始数据的降维处理及显著性表征

原始数据通过线性或非线性方法可以生成任意降维实数张量用于数据检索。新的张量空间的近邻关系可以是任意度量作为相似度，如距离，散度，KID或其他非线性函数生成的正实数。近邻关系度量的唯一需求是保证原始空间相似度的顺序和张量空间相似度度量顺序保持一致。

由于存储空间的限制和个人隐私问题，对原始数据做降维操作并提取实数张量作为显著性表征，数据表征在度量空间中的关系由深度神经网络根据结构化数据标签构建，神经网络训练的损失函数选择将以鼓励相似度度量顺序为主旨。常见损失函数有：单点，配对或列表损失。尽管降维后的数据表征可以选择直接提取神经网络中间层实数张量或对图像多个区域生成多个表征等其他方法，但是降维处理典型并有效的方法是对原始数据做二值化处理操作。哈希算法是指将任意长度的输入向量映射成固定长度的二值码，这个映射称为哈希函数。具体地，图像哈希算法是将原始空间中高维数据映射到低维的二值码空间中，二值码的空间通常远小于原始向量的空间。构造的哈希函数可以是线性或者其他非线性的形式，该函数的设计需要保持二值码空间与原始空间数据的近邻关系(或者相似性)。

哈希算法是一个不可逆的降维并表征的过程，不同的输入可能会形成相同的二值码输出，不能从二值码来确定其唯一的输入。

以深度哈希算法为例，如图A.6所示，使用深度学习模型将对象的图像和标签对应，通过优化哈希目标函数和交叉熵损失函数得到网络参数。然后，将图像输入深度模型，得到二值码。计算该图像二值码与数据库中的图像二值码的汉明距离，并按照汉明距离由小到大返回，得到数据库中相似的图像。



图A.6 深度哈希示意图

A.2.3 差分隐私和特征数据保护

在对图像的二值化特征表示和视频的显著性分析阶段，会存在大量的智能家居训练的隐私数据，对此,提出了一种PATE(Private Aggregation of Teacher Ensembles)差分隐私算法,该算法通过“teacher”和“student”两个模型来降低隐私数据的敏感程度。在传统的隐私保护问题上，使用遮挡唯一认证ID的算法已经不足以满足隐私保护的需求，通过对模型参数或者多类辅助性信息的分析便能够实现差分攻击。PATE算法通过增加噪声的特殊处理对模型产生影响，实现隐私数据“脱敏”处理。算法首先从数据子集中分离出多个数据集并含有隐私数据集，某一个隐私信息只存在于一个分区中。其次将每一个子集训练好的模型聚合，称之为“teacher”模型。所有“teacher”模型形成一个标签后加入噪声，使其达到某一个特定的隐私数据对该类别的选择影响尽可能小。“student”模型从“teacher”模型获取未标签的数据，以隐私保护的方式进行训练。“student”模型从一组未标记的公共数据中选择输入数据，并将输入提交“teacher”模型以获得标签。最后，“student”模型使用标记过的数据来训练模型。该算法对训练数据隐私保护的性能改善主要表现在“teacher”模型的共识度和“student”模型对“teacher”模型的需求程度，标签之前相差越悬殊，性能越好。对“teacher”模型的训练量越少，则对隐私的保护越好，因为“teacher”模型产生标签消耗的隐私预算会被添加到总的成本中。

差分隐私方法可以在原始数据二值化处理的基础上实施，从而实现对隐私数据的保护。

A.2.4 基于通用事件流的隐私保护方法

一种通用的事件流特征提取和编码框架。通过本方法，一方面可以从原始的视频数据中提取出通用的事件流特征，不仅可以支持后续进行多种深度学习应用，也可以进行人眼查验和数据标注，避免数据泄露造成的隐私问题；另一方面，可以对大量的视频数据进行稀疏压缩处理，便于对事件进行传输、存储和分析。

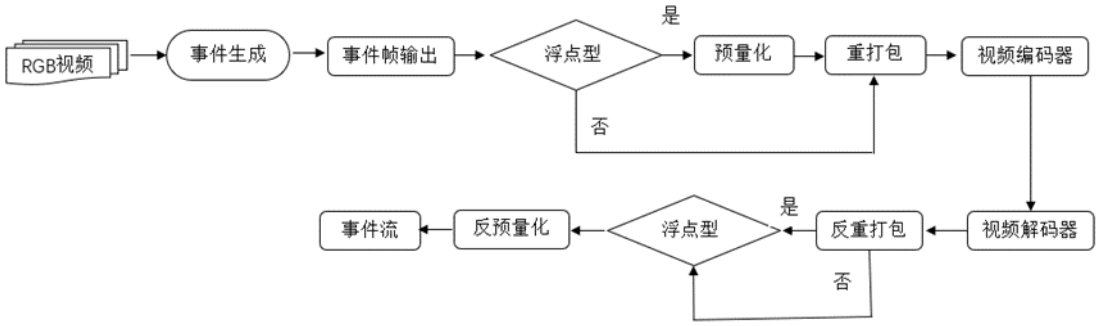
本方案中的通用事件流生成技术框架如图A.7所示，主要包含以下步骤：

a) 事件生成：利用光流法采样生成事件流。首先读取 RGB 源视频，然后使用相应光流算法，生成运动物体的光流数据,包含所有像素点的速度矢量;

b) 事件帧输出：将光流信息进行颜色编码后的输出。首先将每个像素的速度矢量的值进行标准化；然后将标准化后的速度矢量的值使用孟塞尔颜色系统 VHC 编码，得到一张有色图；最后将 VHC 编码转为 RGB 编码，得到最终的事件帧输出；

注：孟塞尔颜色系统VHC，是色度学透过明度(V)、色相(H)、彩度(C)描述颜色的常用方法之一。

c) 视频编解码：利用传统视频编解码器，对事件帧进行编解码。



图A.7 通用的隐私保护事件流生成技术框架