引用格式：邵天睿, 尚涛, 姜亚彤, 等. 基于多分辨率分析的医疗图像大数据差分隐私保护方案 [J]. 电子科技大学学报, 2025, 54(4): 604-617.
SHAO T R, SHANG T, JIANG Y T, et al. Differential privacy protection scheme for medical image big data with multi-resolution analysis[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(4): 604-617.

基于多分辨率分析的医疗图像大数据差分隐私保护方案

邵天睿[1]，尚　涛[1*]，姜亚彤[1]，程宇航[1]，杜瑞忠[2]

（1. 北京航空航天大学 网络空间安全学院，北京 100191；2. 河北大学 网络空间安全与计算机学院，保定 071002）

摘要：医疗图像大数据共享能够有效提升医疗服务质量，为防止医疗图像数据共享过程中发生患者隐私泄露，需要对医疗图像进行隐私保护。差分隐私作为一种具备严格的数学定义和隐私保护强度证明的安全机制，已经应用于图像数据隐私保护。针对医疗图像大数据的隐私保护问题，提出了一种基于多分辨率分析的医疗图像大数据差分隐私保护方案。该方案在 Hadoop 大数据平台上设计图像数据格式，基于 MapReduce 计算框架设计医疗图像大数据差分隐私保护算法；现有图像差分隐私方案对图像中的所有数据进行同等强度的保护，没有考虑不同的数据存在不同的隐私需求，针对此问题，结合医疗图像处理领域常用的小波变换技术，基于小波多分辨率分析提出一种隐私预算分配算法，该算法在小波域内根据各小波频带的隐私需求进行隐私预算分配，根据不同小波频带系数的隐私保护需求进行不同强度的差分隐私保护；最后，设计像素差分扰动算法，基于差分隐私指数机制对图像矩阵中的每个像素进行差分隐私扰动。实验结果表明，该方案能够根据各小波频带的隐私保护需求进行差分隐私保护，且在相同的隐私预算下，该方案的图像视觉效用相比对照方案最多可提升 97.7%，图像分类效用最多可提升 87.2%。在 Hadoop 集群上进行性能测试，该方案能够实现高效的医疗图像大数据差分隐私保护。

关键词：医疗图像；隐私保护；小波变换；差分隐私；大数据
中图分类号：TP309　　**文献标志码：**A　　DOI: 10.12178/1001-0548.2024107

Differential privacy protection scheme for medical image big data with multi-resolution analysis

SHAO Tianrui[1], SHANG Tao[1*], JIANG Yatong[1], CHENG Yuhang[1], and DU Ruizhong[2]

(1. School of Cyber Science and Technology, Beihang University, Beijing 100191, China;
2. School of Cyber Security and Computer, Hebei University, Baoding 071002, China)

Abstract: The release and sharing of medical image big data can effectively improve the quality of medical services. Medical images contain sensitive information of patients, in order to prevent the disclosure of patient privacy during the sharing of medical image data, it is necessary to protect the privacy of medical images. As a security mechanism with strict mathematical definition and proof of privacy protection strength, differential privacy has been widely used in image data privacy protection. To achieve the privacy protection for medical image big data, this paper proposes a differential privacy scheme for medical image big data with wavelet multi-resolution analysis. This scheme designs the medical image data format on the Hadoop platform and designs the differential privacy protection algorithm of medical image big data based on the MapReduce framework. Existing image differential privacy methods protect entire image data with a same privacy level without considering the different privacy requirements of different data. To solve this problem, combining with the wavelet transform technology commonly used in medical image processing, this paper proposes a privacy budget allocation algorithm based on wavelet multi-resolution analysis. The algorithm measures the importance and privacy requirement of different wavelet subbands in wavelet domain, and allocates privacy budget according to the privacy requirement of each wavelet subband. Finally, this paper proposes a pixel differential disturbance algorithm, which disturbs every pixel based on differential privacy exponential mechanism. The experimental results show that the proposed scheme can implement differential privacy protection according to the privacy protection requirements of each wavelet

收稿日期：2024−05−09
基金项目：河北省重点研发计划项目（22340701D）
作者简介：邵天睿，主要从事大数据安全与隐私保护方面的研究。
*通信作者 E-mail: shangtao@buaa.edu.cn

subband. Under the same privacy budget, the image visual effect of this scheme can be improved by up to 97.7% compared with the control scheme, and the image classification effect can be improved by up to 87.2%. The performance experiment on the big data platform shows the proposed scheme can realize efficient differential privacy protection of medical image big data.

Key words: medical image; privacy protection; wavelet transform; differential privacy; big data

医疗影像学为临床提供了 X 射线（X-ray, XR）、电子计算机断层扫描（computed tomography, CT）、磁共振成像（magnetic resonance imaging, MRI）等多种医疗图像信息，医疗机构的诊疗工作越来越依赖于医疗图像的检查。当前我国政府大力推动远程医疗，医疗图像远程共享将为远程诊断提供重要依据。医疗图像包含患者体内各器官的物理信息，属于高度敏感的个人隐私，需要采取隐私保护措施，防止数据共享过程中泄漏患者个人隐私。

经典的图像隐私保护技术有马赛克[1] 和方框模糊[2]，近年来有学者通过生成对抗网络技术[3-4] 对图像进行隐私保护，但这些方法都没有提供严格的隐私保护强度证明。图像加密[5-6] 提供了强大的隐私保护效果，其安全性取决于密钥的安全，一旦密钥泄露将完全丧失隐私保护效果。随着安全多方计算技术的发展，研究人员相继提出了一系列基于同态加密[7]、秘密共享[8-9] 的图像隐私保护方法，这些方案在大数据场景下存在计算开销大、实现难度高的问题[10-11]。差分隐私[12] 是一种新型隐私保护技术，该技术假设攻击者拥有最强大的背景知识，通过对查询结果进行适当的扰动以达到隐私保护的效果，并提供了严格的数学证明。差分隐私在实现隐私保护的同时能够提供较高的数据效用，可以用于解决图像隐私保护问题。

常用的图像处理技术有离散傅立叶变换[13]、离散余弦变换[14] 等，这些技术在计算过程中产生的中间值都是浮点数，大量的浮点数一方面不便于数据存储，另一方面会造成量化误差，而量化误差是图像在处理过程中信息损失最主要的原因。医疗图像对图像质量要求较高，通常要求使用无损处理技术。整数小波变换[15] 在计算过程中产生的中间值都是整形数据，不会产生量化误差，能够对图像实现真正意义的可逆无损处理，广泛应用在医疗图像融合[16]、医疗图像去噪[17] 等场景。小波变换具有方向性特征，人眼对图像小波分解后的各小波频带具有不同的敏感度，不同的小波频带在图像处理中具有不同的重要性，具有不同的隐私保护需求。

医疗图像通常以数字化的方式存储于计算机设备，导致各医疗机构系统内的数字化医疗图像规模迅速增长。据统计，医疗图像数据已经占据医疗数据规模的 80% 以上[18]，医疗图像已然进入大数据时代。个人隐私在大数据时代面临更严峻的风险，亟需探索大数据环境下医疗图像大数据的差分隐私保护方案。

本文的主要贡献包括：1）提出一种 Hadoop 大数据平台上的医疗图像大数据差分隐私保护方案，针对 Hadoop 平台图像处理困难的问题，实现 Hadoop 对图像数据类型及图像处理的支持，基于 MapReduce 计算框架设计方案的算法组件，实现医疗图像大数据的高效隐私保护处理；2）针对医疗图像多分辨率分析场景下，各个小波频带具有不同隐私保护需求的问题，衡量各小波频带的隐私保护需求，根据隐私保护需求强弱为各小波频带进行隐私预算分配，实现依据各小波频带的隐私保护需求进行对应强度的差分隐私保护；3）提出一种像素差分扰动算法，对图像矩阵中的每个元素采用该算法进行差分隐私扰动，在实现医疗图像隐私保护的同时满足医疗图像对图像高质量的要求。

1　相关工作

近年来，图像数据的差分隐私保护得到了研究人员持续关注。文献 [19] 提出了图像差分隐私保护方案，将图像进行马赛克处理，向马赛克图像添加差分隐私噪声实现差分隐私保护。文献 [20-21] 通过奇异值分解提取图像矩阵的特征值，对特征值添加差分隐私噪声，实现对图像的差分隐私保护。文献 [11, 22] 分别采用离散傅立叶变换、离散余弦变换提取图像频域特征值，对特征值添加差分隐私噪声，实现图像的差分隐私保护，类似研究还有文献 [23] 提出的基于基本小波变换的图像差分隐私方案。现有方案大多基于人脸图像展开研究，医疗图像具有明显不同于人脸图像的要求：对图像质量高度敏感，像素之间的细小差异可能代表着不同组织结构或是病理器官的病变程度，轻度的图像失真有可能造成诊断上的严重错误。现有技术方案使用的图像处理技术和差分隐私扰动机制均为浮点数计算，存在量化误差，而量化误差是导致图像在处理过程中出现信息损失的主要原因；另外，这些方案

大多采用舍弃图像细节、提取图像主要特征进行差分隐私保护的技术路线，虽然降低了引入的差分隐私噪声量，但舍弃了部分图像细节信息，难以实际用于医疗图像差分隐私保护。

现有图像差分隐私保护方案对图像中的所有数据施加同等强度的保护，实际上医疗图像中存在前景区域与背景区域，两者对隐私保护的需求存在差异。文献 [24] 提出了一种距离差分隐私机制，其本质是为同一数据集内不同的数据指定不同的隐私预算，从而实现对同一数据集内的数据进行不同强度的差分隐私保护。小波变换是常用的医疗图像处理方法，图像经小波变换得到不同的小波系数，然后在小波域内进行图像处理。不同的小波系数在图像处理中的重要性不同，具有不同的隐私保护需求，目前缺少依据小波频带隐私保护需求进行差分隐私保护的相关研究。

随着医疗图像步入大数据时代，基于 Hadoop 的图像数据处理受到越来越多的关注。Hadoop 不支持直接处理 PNG、JPEG 等常见格式的图像数据，因此首要工作在于设计 Hadoop 支持的图像数据格式。文献 [25-27] 分别设计了图像数据格式，核心思路都是读取图像的二进制流或字节流，利用 Hadoop 支持的 SequenceFile 文件存储图像的流数据，随后重新设计输入输出格式，实现 Hadoop 平台对图像数据在输入、计算、输出、存储等方面的支持。现有的 Hadoop 图像数据格式设计方案存在以下不足：首先，SequenceFile 是 java 语言独有的文件类型，图像数据写入后可读性较差；其次，现有技术方案在图像输入格式的设计中，将每张图像作为一个独立的作业切片，MapReduce 为每个作业切片开启一个 Map 任务进行处理，在图像数量较多时，该处理方法将导致 Map 节点数量过多，频繁开启关闭 Map 节点将对集群性能产生负面影响。

2 预备知识

2.1 整数小波变换

第一代小波变换由定义在$L^2(R)$空间上的容许函数$\psi(t)$经过二进伸缩和平移生成，Mallat 算法[28]的提出使第一代小波变换在图像与信号处理领域得到了广泛应用。第一代小波存在若干局限性，如依赖傅里叶变换构造小波结构，基于卷积运算进行图像处理占用内存较多、计算复杂、小波变换产生大量浮点数、结果不便于数据存储等。

基于提升格式的第二代小波[29] 解决了第一代小波存在的不足，可以实现小波变换的快速简单计算。整数小波变换（integer wavelet transform, IWT）[30] 是一种经典的第二代小波变换，其计算结果是整形数据。新一代图像压缩标准 JPEG 2000[31] 采用整数小波变换实现完全无损的图像压缩，适合医疗图像等对图像质量要求较高的领域。

2.2 小波多分辨率分析

多分辨率分析方法模拟计算机视觉中人眼的感知过程，本质是对图像进行多层分解，在高分辨率下获取图像细节信息，在低分辨率下获取图像轮廓信息，实现对图像在多分辨率下的逐层识别。

小波变换可以将图像逐层分解为低频分量LL_i和水平、垂直、对角线方向的高频分量HL_i、LH_i、HH_i，低频分量LL_i能够以完全相同的方式继续分解为分辨率更小的子图。图像经小波变换分解成若干小波子图后，图像的轮廓信息集中到低频子图，边缘和纹理信息集中到高频子图，达到多分辨率分析的效果。小波多分辨率分析的过程如图 1 所示。

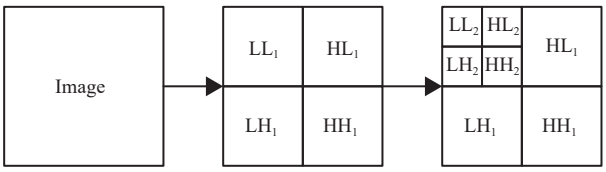


图 1 图像的小波多分辨率分析

本文使用 JPEG 2000 中的整数小波变换，以三级小波分解为例，进行小波多分辨率分析相关算法说明与实验测试。

2.3 差分隐私

差分隐私假设攻击者知晓数据集中除目标个体外的所有其他个体的数据，通过对查询数据按照相应规则施加扰动，实现差分隐私保护。常用的差分隐私扰动机制有拉普拉斯机制和指数机制。

定义 1 拉普拉斯机制[32]。对于给定的查询函数 q 和数据集χ，拉普拉斯机制M_{LAP}的输出为：

$$M_{\mathrm{LAP}}(\chi) = q(\chi) + \mathrm{Lap}(\Delta/\varepsilon) \tag{1}$$

式中，Δ为查询函数 q 的敏感度，ε为单次查询的隐私预算。拉普拉斯机制适用于连续型数据的保护，通过向隐私数据添加服从位置参数为 0、尺度参数为Δ/ε的拉普拉斯分布的噪声，实现对隐私数据的扰动。

定义 2 指数机制[33]。对于效用函数 $f: U\times$

Range $\to R$，存在输入数据集$D \subseteq U$，则效用函数为$U \times$Range中每个(D,r)生成一个效用值$f(D,r)$。指数机制M_{EXP}以正比于$\exp(\varepsilon f(D,r)/2\Delta f)$的概率从Range中选择并输出结果$r$，其中$\Delta f$为效用函数$f$的敏感度。指数机制也是一种差分隐私扰动机制，一般用于离散型数据的保护，表示为：

$$\mathrm{M}_{\mathrm{EXP}}(D,f,\mathrm{Range})=\{r:|\Pr[r\in\mathrm{Range}]\propto\exp(\varepsilon f(D,r)/2\Delta f)\} \tag{2}$$

Range为有限集合或无限集合时，指数机制均适用，由于需要对所有的(D,r)计算效用值，所以当Range为有限集合时指数机制的应用效果更好。

定理 1　串行组合性[32]。假设存在n个随机机制$M_1,M_2,\cdots,M_n$，每个机制$M_i(i\in[1,n])$各提供ε_i-差分隐私。对于数据集D，由n个机制依次构成的组合机制$M(D)=(\mathrm{M}_1(D),\mathrm{M}_2(D),\cdots,M_n(D))$满足$\varepsilon-$差分隐私，其中：

$$\varepsilon=\sum_{i=1}^{n}\varepsilon_i$$

定理 2　并行组合性[34]。假设存在n个随机机制$M_1,M_2,\cdots,M_n$，每个机制$M_i(i\in[1,n])$分别提供ε_i-差分隐私。若每个机制M_i分别作用于数据集D的互不相交的子集$D_i(i\in[1,n])$，则组合机制满足ε-差分隐私，其中：

$$\varepsilon=\max(\varepsilon_i)$$

定理 3　后处理免疫[35]。对于满足$\varepsilon-$差分隐私的随机机制$M:N^{|x|}\to R$，其中$N^{|x|}$为随机机制M的定义域。令$f:R\to R'$是任意的随机映射，则$f\cdot M:N^{|x|}\to R'$仍满足$\varepsilon-$差分隐私。

2.4　HDFS 与 MapReducce

HDFS 是 Hadoop[36] 中负责管理大数据存储的分布式文件系统，采用主从架构，由一个管理节点 NameNode 和多个工作节点 DataNode 组成，NameNode 负责管理命名空间、存储元数据，DataNode 负责实际存储数据。

MapReduce 是一种面向大数据的分布式并行编程模型和计算框架，MapReduce 将计算任务抽象成 Map、Reduce 以及可选的 Combine 阶段，每个阶段都将数据以键值对的形式输入和输出。Map 负责将输入数据根据用户定义的业务逻辑进行并行计算；Combine 在 Map 端对 Map 输出的中间数据进行预聚合，属于一种优化机制，并非所有程序都需要此阶段；Reduce 拉取 Map 输出的数据，执行用户业务逻辑，对数据按照键的字典序进行排序，合并具有相同键的值，最终将计算结果输出到 HDFS 中。

3　基于多分辨率分析的医疗图像大数据差分隐私保护方案

本节介绍基于多分辨率分析的医疗图像大数据差分隐私保护方案（medical image big data differential privacy scheme with wavelet multi-resolution analysis, MIBD2P），该方案针对小波域内医疗图像不同区域不同的隐私需求进行隐私预算分配，设计图像的差分隐私扰动算法，在大数据平台对医疗图像进行隐私保护。具体地，MIBD2P 方案由基于多分辨率分析的医疗图像大数据差分隐私保护算法（multi-resolution analysis based medical image big data differential privacy algorithm, MRADP）、基于小波多分辨率分析的隐私预算分配算法（wavelet multi-resolution analysis based budget allocate algorithm, WMABA）以及像素差分扰动算法（pixel differential disturbance algorithm, PD2）组成。

3.1　基于多分辨率分析的医疗图像大数据差分隐私保护算法

由于 Hadoop 不支持直接处理图像数据，因此本小节首先设计 Hadoop 平台的图像存储与输入输出格式，并在此数据格式的基础上进行图像差分隐私保护算法的设计。

3.1.1　Hadoop 图像存储数据格式设计

TextFile 即文本文件，是 Hadoop 默认的文件格式，各种编程语言、系统平台均支持读写文本文件，可读性、扩展性强，使用 TextFile 存储数据具有 SequenceFile 不具备的优势。本文使用 TextFile 存储图像数据，使用 Text 数据类型构造 Hadoop 图像数据格式。

在对图像进行差分隐私处理的过程中，只需对图像的像素值进行扰动，图像的元数据信息在处理前后保持一致。因此，可将图像的路径、名称等元数据信息作为图像在 MapReduce 处理过程中的 Key。Java 的 BufferedImage 工具类能够将图像数据读取为字节流，将字节流进行 Base64 编码即可将图像数据转为字符串在文本文件中进行存储，可以借助 BufferedImage 进行图像 Value 的设计。综上，本文设计 Hadoop 图像数据的存储格式如下：1）Key：各图像在文件系统的绝对路径；2）Value：编码的 BufferedImage 字符串。

在 Hadoop 平台存储图像数据前，将每张图像的键值对信息以 Key+“\t”+Value 格式拼接为一个字符串，再存储为文本数据（为便于描述，后续将这种表示图像键值对信息的一行文本，统一简称为“图像文本”）。

3.1.2 Hadoop 图像输入输出数据格式设计

InputFormat 是 MapReduce 计算流程中的第一个组件，规定了在 Hadoop 中如何读取和拆分输入文件，并将文件切片分配给 Mapper 程序进行处理。OutputFormat 则是 MapReduce 流程中的最后一个组件，规定了 Reduce 的计算结果应当如何输出。为了将图像转换为图像文本进行处理，并在处理结束后进行图像文本的存储，需要实现自定义输入格式 InputFormat 和输出格式 OutputFormat。本文根据医疗图像大数据差分隐私保护算法的需求，针对现有研究的不足，设计实现了以下两个类。

1）ImageTextInputFormat。继承自基类 FileInputFormat，主要具有 2 个功能：将输入路径下的所有图像按照 3.1.1 中图像文本的格式集中存储在 HDFS 上的一个文本文件中，文本文件的每一行都是一张图片的键值对信息；根据终端输入的参数控制每个 Map 节点处理的图像数量，解决现有图像输入格式设计方案将每张图片均作为一个独立切片的不足，具有在启动 MapReduce 任务时通过终端参数直接指定 Mapper 程序并行度的功能。

2）ImageTextOutputFormat。继承自基类 FileOutputFormat，主要具备 2 个功能：将 Reduce 阶段输出的 key 和 Value 进行编码，并转换为一行图像文本；将图像文本写回到 HDFS 的指定目录，并仍将数据集中所有的图像信息储存在一个文本文件中，文本文件的每一行代表一张差分隐私保护后的图像。

3.1.3 基于 MapReduce 框架的 MRADP 算法设计

基于 MapReduce 框架设计 MRADP 算法，包含 Mapper、Combiner、Reducer 这 3 个子算法。Mapper 程序接收来自 ImageTextInputFormat 的文件切片，对切片中的每个图像进行整数小波变换、隐私预算分配和差分隐私处理，并将差分隐私保护后的各个小波子频带的键值对输出到 MapReduce 上下文。Mapper 程序的实现步骤如算法 1 所示。

算法 1 MRADP-Mapper

输入：图像 $\boldsymbol{A}$ 的图像文本$\text{ImageText}_{\boldsymbol{A}}$，隐私预算$\varepsilon$

输出：图像 $\boldsymbol{A}$ 经过差分隐私扰动的各个小波子频带的键值对$(\text{Key}_{\boldsymbol{A}},\text{Value}_{\boldsymbol{A}})$ ←解码$\text{ImageText}_{\boldsymbol{A}}$获取图像 $\boldsymbol{A}$ 的键值对；

$\boldsymbol{A}$ ←解码$\text{Value}_{\boldsymbol{A}}$获取图像 $\boldsymbol{A}$ 的像素矩阵；

subbands ←将矩阵$\boldsymbol{A}$进行三级小波分解获得所有小波频带；

ε_list ← WMABA($\boldsymbol{A},\varepsilon$)；//为小波频带分配隐私预算

For 各频带subband与其隐私预算ε_i Do：

For subband中的各小波系数pixel Do：

$\text{pixel}' \leftarrow \text{PD2}(\text{pixel},\varepsilon_i)$；

End For

$\text{subband}'$ ←将扰动后的小波系数组成扰动后的小波频带；

为该小波频带设置隐私需求标识 index；

value ←将 index 和$\text{subband}'$拼接转换为字符串；

将$(\text{Key}_{\boldsymbol{A}},\text{value})$写入 MapReduce 上下文；

End For

Mapper 程序将每一行图像文本 $\text{ImageText}_{\boldsymbol{A}}$ 切分为键$\text{Key}_{\boldsymbol{A}}$、Base64 编码的 BufferedImage 字符串$\text{Value}_{\boldsymbol{A}}$，并进一步将$\text{Value}_{\boldsymbol{A}}$解码恢复 BufferedImage 类型的图像 $\boldsymbol{A}$。通过 WMABA 算法为$\boldsymbol{A}$的所有子频带分配隐私预算，使用 PD2 算法分别对每个子频带进行差分隐私保护，最后将差分隐私保护后的子频带的键值对写入 MapReduce 上下文。WMABA 算法根据小波频带的隐私保护需求进行隐私预算分配，因此设置字段 index 用以标识各小波频带的重要性，即隐私需求。关于 Mapper 程序输出键值对的设计，由于一个 Mapper 输出的所有小波频带均来自同一张图像 $\boldsymbol{A}$，所以将各小波频带的键均设置为原图像 $\boldsymbol{A}$ 的键；最后将小波频带的 index、小波系数统一存储在一个字符串中作为该小波频带的值 value。

在 Hadoop 中，小文件指的是文件大小远小于 HDFS 文件块大小的文件，Hadoop 存在小文件存储瓶颈问题。为避免在 HDFS 上存储大量的小文件，本文将 MapReduce 处理后的全部图像写入同一个文本文件中进行存储，因此 MRADP 只能有一个 Reducer 程序。同时，在三级小波变换的案例下，MRADP-Mapper 程序输出的中间数据量将是输入数据量的 10 倍。若待处理的图像数量较多，Mapper 程序产生的大量中间临时数据交由唯一的 Reduce 节点进行数据汇聚，存在诸多弊端：一方面将增大 Map 节点与 Reduce 节点之间的数据传输量，占用大量集群带宽，另一方面可能导致 Reduce 节点负担过重。因此本文引入 Combiner 程序，在

Map 端对 Mapper 程序输出的键值对数据进行预聚合，提升程序汇聚阶段的并行度。Combiner 程序的实现步骤如算法 2 所示。

算法 2　MRADP-Combiner

输入：由同一图像 $\boldsymbol{A}$ 分解得到的所有小波频带的键值对$(\mathrm{Key}_{\boldsymbol{A}},\mathrm{value})$

输出：差分隐私保护后的图像$\boldsymbol{A}'$的键值对

For 每个小波频带键值对$(\mathrm{Key}_{\boldsymbol{A}},\mathrm{value})$ Do：

$\mathrm{index},\mathrm{subband}' \leftarrow$解码 value；

End For

$(\mathrm{LL}_3', \mathrm{HL}_3', \mathrm{LH}_3', \mathrm{HH}_3', \mathrm{HL}_2', \mathrm{LH}_2', \mathrm{HH}_2', \mathrm{HL}_1', \mathrm{LH}_1', \mathrm{HH}_1')\leftarrow$将所有的$\mathrm{subband}'$根据 index 标识进行识别；

$\mathrm{LL}_2' \leftarrow$将$\mathrm{LL}_3', \mathrm{HL}_3', \mathrm{LH}_3', \mathrm{HH}_3'$进行逆小波变换；

$\mathrm{LL}_1' \leftarrow$将$\mathrm{LL}_2', \mathrm{HL}_2', \mathrm{LH}_2', \mathrm{HH}_2'$进行逆小波变换；

$\boldsymbol{A}' \leftarrow$将$\mathrm{LL}_1', \mathrm{HL}_1', \mathrm{LH}_1', \mathrm{HH}_1'$进行逆小波变换；

将$\boldsymbol{A}'$转换为字符串，作为图像$\boldsymbol{A}'$的值value'；

将$(\mathrm{Key}_{\boldsymbol{A}},\mathrm{value}')$写入 MapReduce 上下文.

一个 Mapper 程序输出的所有键值对都会被同一个 Combiner 程序接收，但无法保证接收的顺序。逆小波变换的输入需严格依照各频带的层次和方向顺序，因此 Combiner 程序借助 index 字段进行小波频带的识别，将乱序的各$\mathrm{subband}'$进行排序。由于 Mapper 程序对输入图像进行了三级小波分解，因此 Combiner 程序进行了 3 次逆变换，最终恢复原分辨率的图像$\boldsymbol{A}'$。Combiner 程序完成了小波频带的预聚合以及逆小波变换处理，得到了与原图像 $\boldsymbol{A}$ 分辨率相同的扰动图像。Reducer 程序只需拉取 MapReduce 上下文中的键值对并传递给 ImageTextOutputFormat 输出即可，在很大程度上减轻了 Reduce 节点的计算负担。Reducer 程序的实现步骤如算法 3 所示。

算法 3　MRADP-Reducer

输入：具有相同 Key 的图像 Value 集合

输出：集合中所有的键值对

读取 MapReduce 上下文中的键值对数据；

将读取的键值对数据写入 MapReduce 上下文.

MRADP 算法的驱动程序负责 configuration 对象、job 任务等一系列关键运行参数的配置工作，包括输入类与输出类、输入键值对类型与输出键值对类型、输入图像路径与输出图像文本路径等，配置完成后启动 MapReduce 任务。

3.2　基于小波多分辨率分析的隐私预算分配算法

小波频带隐私保护需求如图 2 所示，图中箭头方向为小波频带隐私保护需求逐渐降低的方向。

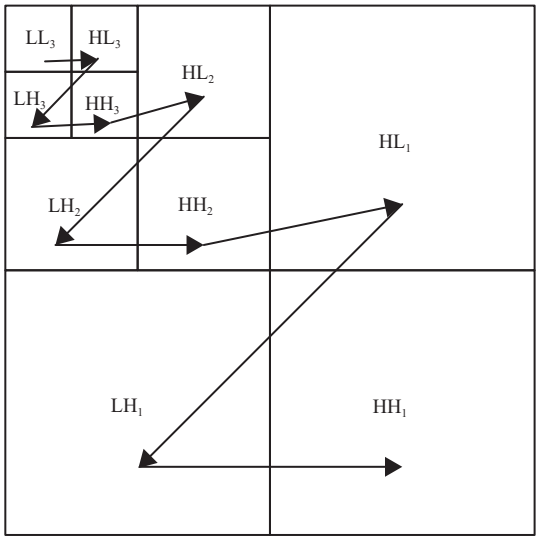


图 2　小波频带隐私保护需求

小波多分辨率分析是小波图像处理中必不可少的分析方法，医疗图像需要在多分辨率下进行图像配准、图像降噪等处理。EZW[37]、SPIHT[38]、EBCOT[39]是 3 种经典的基于多分辨率分析的小波图像编码算法。EZW 算法将图像小波分解后，基于低频子带构造零树结构，按照水平高频、垂直高频、对角线高频的顺序依次进行扫描，并使用低频子带系数对高频子带系数进行编码；SPIHT 算法基于低频子带构造空间方向树，分别从水平、垂直、对角线 3 个方向依照低分辨率到高分辨率的顺序对小波系数进行编码；EBCOT 是 JPEG 2000 的核心编码算法，其编码过程从最左上角的小波系数开始，至右下角的小波系数结束，其在图像码流的压缩过程中也对右下角的小波系数进行选择性截断。

根据经典小波图像编码算法原理，可知小波系数的重要性与以下因素有关：1）小波系数幅值的大小：幅值大的小波系数比幅值小的重要；2）小波系数所在的子频带：低分辨率子频带的小波系数要比高分辨率子频带的小波系数重要；3）人眼的视觉敏感程度：根据人类视觉系统的理论模型，人眼对低频系数比对高频系数敏感，对垂直和水平方向子图小波系数比对角线方向子图小波系数敏感。低频子图较小的小波系数往往比高频子图较大的小波系数具有更重要的视觉意义，水平和垂直方向子图较小的小波系数往往也比对角线方向子图较大的小波系数更重要。

数据越重要，隐私保护需求越高，综上分析可得多分辨率分析场景下各小波频带系数的重要性与隐私保护需求规律：不同分辨率子频带的小波系数之间，小波频带分辨率越低，小波系数越重要，隐私保护需求越高；同一分辨率的各个子频带之间，

重要性与隐私保护需求由高到低依次为低频子带、水平高频子带、垂直高频子带、对角高频子带。

在差分隐私中，隐私预算的大小表征隐私保护的强弱，本文根据各小波频带不同的隐私保护需求，提出 WMABA 算法。WMABA 算法的流程如算法 4 所示。

算法 4　WMABA 算法

输入：隐私预算ε，图像$\boldsymbol{A}$

输出：图像$\boldsymbol{A}$各个小波子频带分配的隐私预算

$(\mathrm{LL}_1,\mathrm{HL}_1,\mathrm{LH}_1,\mathrm{HH}_1)\leftarrow$对$\boldsymbol{A}$进行小波分解；

$(\mathrm{LL}_2,\mathrm{HL}_2,\mathrm{LH}_2,\mathrm{HH}_2)\leftarrow$对低频带$\mathrm{LL}_1$进行小波分解；

$(\mathrm{LL}_3,\mathrm{HL}_3,\mathrm{LH}_3,\mathrm{HH}_3)\leftarrow$对低频带$\mathrm{LL}_2$进行小波分解；

subbands ←得到图像$\boldsymbol{A}$经三级小波分解的所有子频带；

For 每个小波频带subband Do：

　计算subband携带的信号能量E_i；

End For

$E\leftarrow\sum E_i$；

$\rho\leftarrow E_{\mathrm{LL}_3}/E$；

$\varepsilon_m\leftarrow(1-\rho)\varepsilon$，$\varepsilon_M\leftarrow\varepsilon$；

ε_list ←采用等差数列的方式，对其他小波子频带按照隐私保护需求由高到低的顺序，在$(\varepsilon_m,\varepsilon_M)$区间内分配隐私预算；

Return ε_list.

整数小波变换具有正交性，原图像经小波变换后产生的子频带之间相互独立，满足定理 2 中并行数据集的特征，因此在隐私预算的分配过程中适用差分隐私并行组合定理。WMABA 首先确定隐私保护需求最高的低频子带LL_3的隐私预算，通过计算其携带的信号能量占全部小波频带总信号能量的比例ρ，量化评估其在全部小波频带中的重要性。差分隐私的隐私预算大小和隐私保护强度为负相关关系，因此计算低频子带LL_3的隐私预算为$\varepsilon_m=(1-\rho)\cdot\varepsilon$；由于整张图像需要满足$\varepsilon$−差分隐私，所以为隐私需求最低的频带$\mathrm{HH}_1$分配预算$\varepsilon_M=\varepsilon$；其他小波频带的重要性介于$\mathrm{LL}_3$与$\mathrm{HH}_1$之间，按照隐私需求由强到弱的顺序采用等差的方式为其他频带在$(\varepsilon_m,\varepsilon_M)$区间内分配隐私预算。差分隐私的并行组合[34]以及后处理免疫[35]性质保证最终恢复的图像满足ε−差分隐私。

3.3　像素差分扰动算法

在给定效用函数f时，差分隐私指数机制评估每个可能的扰动结果的数据效用，并依概率返回数据效用最高的数值作为差分隐私扰动结果。指数机制在实现差分隐私的同时，往往能较好地兼顾输出数据的可用性，因此常用于设计启发式算法选取最优参数[21-22, 40]。

医疗图像多是单通道灰度图，其像素取值为[0, 255]范围内的整数。图像经整数小波变换后，生成的小波系数的数值依然是有限范围的整数，因此可以通过指数机制构造图像像素的差分隐私扰动方法。本文基于指数机制提出 PD2 算法，实现图像矩阵中单个像素的差分隐私保护，输出差分隐私扰动后的像素值。PD2 算法中的相关定义如下。

定义 4　查询函数。查询函数$Q_{\boldsymbol{A}}$：输出图像矩阵$\boldsymbol{A}$中某个元素A_{mn}的数值。

定义 5　效用函数。为提升数据效用，在实现差分隐私保护的基础上，使得查询函数的输出尽可能地接近扰动前的原数据，设计效用函数f：

$$f=|\mathrm{Range}|-\mathrm{abs}(A_{mn}-r) \tag{3}$$

式中，Range为元素A_{mn}的取值范围，即查询函数$Q_{\boldsymbol{A}}$的值域；|Range|为Range内的元素数量；abs 代表绝对值计算；r为Range中的任意数值。效用函数f为每一个可能的r计算其效用值，效用值越大的r值，被查询函数输出的概率越高。

在效用函数f的控制下，PD2 算法将在满足差分隐私的同时，尽可能地输出与原数据相差最少的扰动值，从而使扰动后的数据拥有较高的数据效用。PD2 算法的过程如算法 5 所示。

算法 5　PD2 算法

输入：待保护的像素A_{mn}，A_{mn}所在图像矩阵$\boldsymbol{A}$中的最大值Max与最小值min，隐私预算ε

输出：差分隐私扰动后的像素A'_{mn}

Range ← [min,Max]，|Range| = Max − min +1；

For Range集合中的每个r值 Do：

　$f(r)\leftarrow|\mathrm{Range}|-\mathrm{abs}(A_{mn}-r)$；

　$w_r\leftarrow\exp(\varepsilon f(r)/2\Delta f)$；

End For

Sum $\leftarrow\sum w_r$；

For Range集合中的每个r值 Do：

　$\Pr(r)\leftarrow w_r/\mathrm{Sum}$；

End For

依概率$\Pr(r)$将A_{mn}扰动为$A'_{mn}=r$；

Return A'_{mn}.

PD2 算法首先根据最大值 Max 与最小值 min 确定像素A_{mn}及其差分隐私扰动结果A'_{mn}的取值范围Range。首先，对像素A_{mn}在Range内所有可能的

取值 r 按照式 (3) 进行效用评估，接着，依据效用评估值计算每个 r 值的权重，然后，依据权重计算像素 A_{mn} 被扰动为每个 r 的概率。权重越高的数值，其输出概率越大，最终满足差分隐私的概率输出扰动值。

PD2 算法实现对单个像素的 ε–差分隐私扰动，该算法满足差分隐私的证明如下。

对于像素矩阵 $\boldsymbol{A}$ 中的一个像素 A_{mn}，定义其邻近数据 B_{mn} 为与 A_{mn} 相差为 1 的另一个像素数值。对像素 A_{mn} 和其临近数据 B_{mn}，推导如下：

$$\frac{\Pr[\mathrm{M}_{\mathrm{PD2}}(A_{mn},f,\mathrm{Range})=r]}{\Pr[\mathrm{M}_{\mathrm{PD2}}(B_{mn},\mathrm{f},\mathrm{Range})=r]}=\frac{\exp\left(\frac{\varepsilon f(A_{mn},r)}{2\Delta f}\right)}{\frac{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}{\exp\left(\frac{\varepsilon f(B_{mn},r)}{2\Delta f}\right)}}=\frac{\exp\left(\frac{\varepsilon f(A_{mn},r)}{2\Delta f}\right)}{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(B_{mn},r')}{2\Delta f}\right)}$$

$$\frac{\exp\left(\frac{\varepsilon f(A_{mn},r)}{2\Delta f}\right)}{\exp\left(\frac{\varepsilon f(B_{mn},r)}{2\Delta f}\right)}\frac{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(B_{mn},r')}{2\Delta f}\right)}{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}=$$

$$\exp\left(\frac{\varepsilon\Delta f}{2\Delta f}\right)\frac{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(B_{mn},r')}{2\Delta f}\right)}{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}\leqslant$$

$$\exp\left(\frac{\varepsilon}{2}\right)\frac{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')+\Delta f}{2\Delta f}\right)}{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}\leqslant$$

$$\exp\left(\frac{\varepsilon}{2}\right)\exp\left(\frac{\varepsilon}{2}\right)\frac{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}{\sum_{r'\in\mathrm{Range}}\exp\left(\frac{\varepsilon f(A_{mn},r')}{2\Delta f}\right)}=\exp(\varepsilon)$$

证毕。

由定理 2 可知，使用 PD2 算法对图像矩阵内所有像素值同时进行差分隐私处理，得到的图像依然满足 ε–差分隐私。

4　实验分析

实验采用 MedNIST 医疗图像数据集[41] 和 Alzheimer 神经影像数据集[42]，MedNIST 数据集包括头部 CT、手部 CT、胸部 CT、腹部 CT、胸部 XR 和乳腺 MRI 这 6 类共约 60 000 张 64×64 大小的医疗图像，Alzheimer 数据集包括正常、轻度病症、非常轻度病症共 619 张 176×208 大小的医疗 MRI 图像。4.1～4.3 节实验所用的编程语言为 Python 3.9，在 PyCharm 集成开发环境中编写和调试，操作系统为 Windows 11，CPU 为 Intel(R) Core (TM) i5-11 400 @ 2.70GHz；4.4 节实验使用的编程语言为 Java 1.8，运行环境为 IntelliJ IDEA，操作系统为 deepin 20.9，CPU 为 Intel(R) Core(TM) i5-11 400 @ 2.70 GHz。实验数据为 10 次取平均值。

4.1　隐私预算分配测试

在总的隐私预算 ε 为 1.0 时，测试 WMABA 算法的隐私预算分配功能。在 MedNIST 数据集的 6 类图像中分别选择一张图像，对各图像分别进行三级整数小波分解，计算分解后的各小波子频带的能量占所有频带总能量的比例，如表 1 所示。由表 1 可见，图像三级小波分解后的低频带 LL_3 总是携带有最多的图像能量，说明 LL_3 是最重要的子频带，与理论分析结果一致，应当对其进行最严格的隐私保护。根据 WMABA 算法为各小波频带分配隐私预算，隐私预算分配结果如表 2 所示，其总预算为 1.0。

对照观察表 1 与表 2 中的数据，低频子带 LL_3 能量占比越高的图像，为 LL_3 分配的隐私预算越少。LL_3 能量占比越高，表明其携带的图像的主要信息越多、隐私保护需求越高，应当为其分配更少的隐私预算以实现更强的隐私保护。其他小波频带所分配的隐私预算则依据其重要性的降低而以等差数列的方式逐渐增大。因此 WMABA 算法能够实现依据小波频带的隐私保护需求，施加不同强度的差分隐私保护的功能。

表 1　三级小波分解后不同子频带的能量分布　　%

测试医疗图像	三级小波分解后的各子频带									
	LL_3	HL_3	LH_3	HH_3	HL_2	LH_2	HH_2	HL_1	LH_1	HH_1
头部CT	19.6	5.4	4.4	4.6	11.2	6.3	11.4	16.7	10.1	10.3
手部CT	36.1	11.8	2.4	3.5	13.7	6.6	2.6	17.4	3.8	2.1
胸部CT	54.4	0.5	0.3	1.1	1.2	2.5	4.2	6.8	8.6	20.3
腹部CT	41.2	2.4	2.7	5.7	6.4	9.1	9.6	9.8	9.3	3.8
胸部XR	52.0	7.7	5.1	2.4	6.3	11.3	3.1	6.1	4.5	1.5
乳腺MRI	16.3	5.9	5.7	10.1	14.0	8.4	9.4	12.0	14.3	4.0

表 2 不同子频带分配的隐私预算值

测试医疗图像	三级小波分解后的各子频带									
	LL_3	HL_3	LH_3	HH_3	HL_2	LH_2	HH_2	HL_1	LH_1	HH_1
头部CT	0.804	0.825	0.847	0.869	0.891	0.913	0.935	0.956	0.978	1.0
手部CT	0.639	0.679	0.719	0.759	0.799	0.839	0.880	0.920	0.960	1.0
胸部CT	0.456	0.516	0.577	0.637	0.698	0.758	0.819	0.879	0.940	1.0
腹部CT	0.588	0.634	0.680	0.726	0.771	0.817	0.863	0.909	0.954	1.0
胸部XR	0.479	0.537	0.595	0.653	0.711	0.769	0.826	0.884	0.942	1.0
乳腺MRI	0.837	0.855	0.873	0.891	0.910	0.928	0.946	0.964	0.982	1.0

4.2 图像视觉安全评估

图像视觉安全评估分为主观评估和客观评估，用于衡量处理后的图像对于人眼的不可理解程度。图像的视觉安全性越高，非授权用户能从中获得的原图像的信息越少。

本节在 MedNIST 数据集中选取一张胸部 XR 图片，采用本文 MIBD2P 方案、Pix 方案[18]、SRA 方案[20]、QAPP 方案[11] 和 WIP 方案[23] 在不同的隐私预算下分别对该图像进行差分隐私处理，评估处理后的图像视觉安全性。

4.2.1 主观视觉评估

图 3 展示了在不同的隐私预算下，5 种图像差分隐私保护方案处理后的图像，观察各图像的视觉效果。

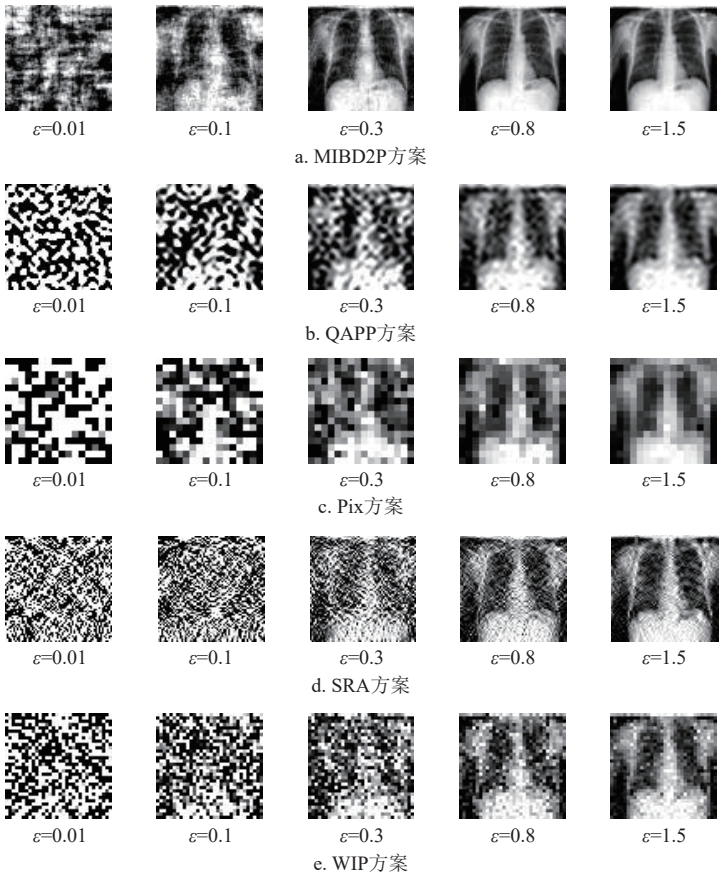


图 3 各差分隐私保护方案扰动后的胸部 XR 图像

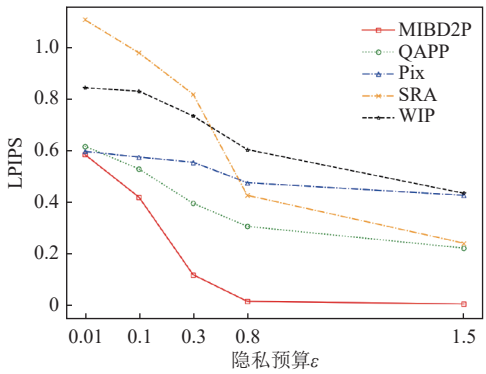
由图 3 可见，随着隐私预算的增大，各方案输出的差分隐私保护的图像均逐渐清晰，这是因为随着隐私预算的增大，差分隐私的扰动幅度变小。在相同的隐私预算下，MIBD2P 扰动后的图像清晰度总是高于其他对照方案。

4.2.2 客观视觉评估

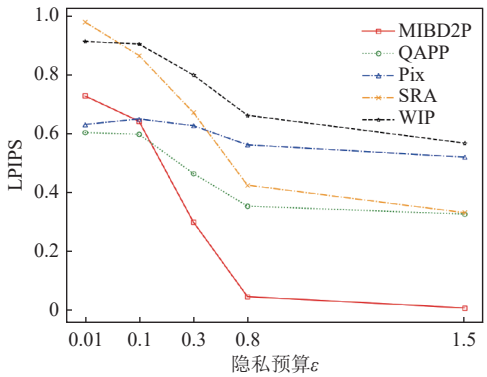
学习感知图像块相似度[43]（learned perceptual image patch similarity, LPIPS）用于度量 2 张图像之间的差别，该方法对深度学习网络抽取的特征进行比对，是一种定量评估方法。LPIPS 认为，即使 2 个图像在像素级别上非常接近，人类观察者也可能将它们视为不同。LPIPS 使用预训练的深度网络（如 VGG、AlexNet）来提取图像特征，然后计算这些特征之间的距离，以评估图像之间的感知相似

度。LPIPS 比传统的结构相似度、峰值信噪比等衡量指标更符合人眼的感知情况，其数值越低，表示 2 张图像越相似。实验选取 MedNIST 和 Alzheimer 数据集中的图像，通过计算不同隐私预算下各方案扰动后的图片与原图的 LPIPS 值，对图像的视觉效果进行客观评估，结果如图 4 所示。

观察图 4，当隐私预算较小时（$\varepsilon \leqslant 0.1$），差分隐私对原图像数据的扰动幅度较大，各方案在 MedNIST 数据集和 Alzheimer 数据集上的 LPIPS 数值均处于较高水平，意味着各方案输出的图像与原图像均存在较大差别。随着隐私预算的增加，本文方案输出图像的 LPIPS 数值迅速降低，该现象有以下几方面的原因：首先，MIBD2P 方案使用基于指数机制设计的 PD2 算法进行差分隐私扰动，该算法输出高效用数据的概率随着隐私预算的增大会迅速增大，因此本方案输出的隐私保护图像与原图像的 LPIPS 值会随着隐私预算的增大而迅速下降；其次，QAPP、WIP 在图像处理过程中舍弃了原图像的部分细节信息，Pix 的马赛克处理同样造成了原图像细节信息的丢失，MIBD2P 没有舍弃图像细节信息；最后，浮点数计算使得各对照方案在图像处理过程中存在由浮点数的量化误差导致的额外信息损失，MIBD2P 使用完全可逆的整数小波变换处理图像，不存在量化误差。



a. MedNIST数据集



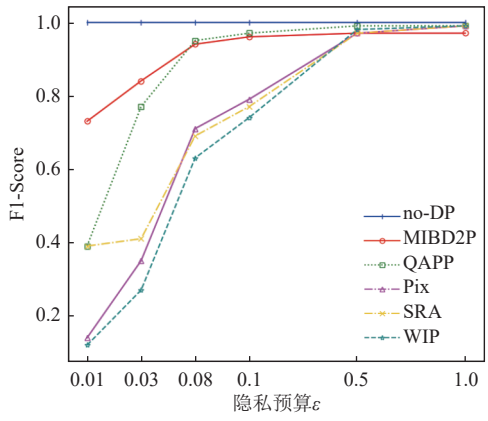
b. Alzheimer数据集

图 4　不同数据集上各方案处理后的图像 LPIPS 对比

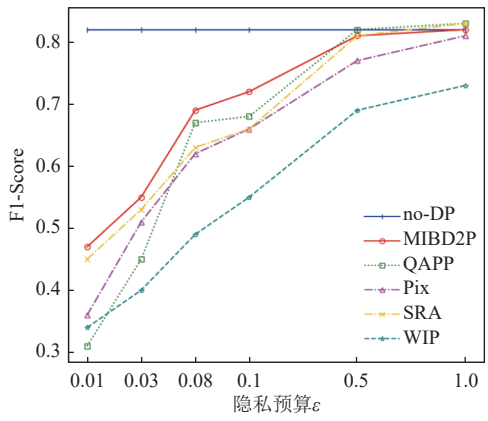
综合图像主观视觉、客观视觉评估实验结果可知，MIBD2P 方案处理后的图像在满足差分隐私保护的同时，相比其他方案具有更高的视觉效用。进一步比较整体效果更好的 QAPP 和 MIBD2P 方案，在 MedNIST 数据集上，当$\varepsilon \geqslant 0.1$时，MIBD2P 方案的图像视觉效果相比 QAPP 方案能够提升 20.7%～97.7%；在 Alzheimer 数据集上，当$\varepsilon \geqslant 0.3$时，MIBD2P 方案的图像视觉效果相比 QAPP 方案能够提升 35.3%～97.5%。

4.3　图像分类效用评估

实验采用的图像分类方法是支持向量机（support vector machine, SVM），MedNIST 数据集在每类图像中选取前 100 张作为原数据集，Alzheimer 数据集使用全部图像作为原数据集。MIBD2P、Pix、QAPP、SRA 和 WIP 各方案在不同的隐私预算下分别处理图像，生成差分隐私处理后的扰动数据集。对于每一个扰动数据集，选取其中 50% 的图像为训练集，50% 的图像为测试集进行 SVM 模型训练。实验以 F1-Score 指标来验证图像分类的效果，5 种方案的 SVM 分类结果如图 5 所示。



a. MedNIST数据集



b. Alzheimer数据集

图 5　不同数据集上各方案图像分类 F1-Score 对比

随着隐私预算从 0.01 增加到 1.0，Pix、QAPP、SRA、WIP 与 MIBD2P 各方案的 F1-Score 值均在增加，原因是差分隐私扰动幅度随着隐私预算的增加而减小，各方案处理后的图像均逐渐接近原图像，图像分类效果不断提升。在隐私预算$\varepsilon \geqslant 0.5$时，各方案在 MedNIST 数据集与 Alzheimer 数据集的 F1-Score 均已接近原数据集。当隐私预算较小时（$\varepsilon \leqslant 0.5$），各方案在 MedNIST 数据集上的分类效果差距较大，MIBD2P 方案和 QAPP 方案相比 Pix、SRA、WIP 方案仍然具有较高的 F1-Score，说明 MIBD2P 和 QAPP 方案在 MedNIST 数据集上的效果整体优于其他方案；当$\varepsilon \leqslant 0.08$时，MIBD2P 方案的图像分类效果开始优于 QAPP 方案，当$\varepsilon = 0.01$时，MIBD2P 方案的 F1-Score 比 QAPP 方案高约 87.2%。在 Alzheimer 数据集上，各方案的图像分类效果整体差距较小，但在较小的隐私预算下 MIBD2P 方案依然拥有最好的图像分类效果，当$\varepsilon \leqslant 0.1$时，MIBD2P 方案的 F1-Score 相比各对照方案高约 3.0%～5.9%；当$\varepsilon = 1.0$时，QAPP、SRA、MIBD2P 方案均出现了 F1-Score 略高于未加噪数据集 F1-Score 的情况，由于未加噪 Alzheimer 数据集上的图像分类结果本身较低，噪声的引入可能导致某些图像特征相比原图得到增强，因此在一些分类效果较好的方案上出现了分类结果高于未加噪数据集的情况，通过增加实验次数应当可以消除这种随机性。

综合不同数据集上的实验分析结果，本文提出的图像差分隐私保护方案在提供ε–差分隐私保护的同时，相比 Pix、QAPP、SRA、WIP 方案具有更高的图像分类效用。

4.4 大数据性能评估

实验在单台物理机和 VMware 虚拟机进行，物理机部署 NameNode 和 1 个 DataNode 节点，再由 VMware 搭建 2 台虚拟机作为 DataNode 节点。物理机配置为 6 核 12 线程处理器，16 G 内存，deepin 20.9 操作系统，Hadoop 版本为 2.7.7；虚拟机分配 2 核 2 线程处理器，4 G 内存，操作系统为 Centos 7，Hadoop 版本为 2.7.7，各节点均采用 Hadoop 默认配置。

在 MedNIST 数据集和 Alzheimer 数据集上，采用加速比 S、可扩展性 K 两项指标[44] 对程序性能进行验证。S 用于衡量在增加 Map 节点时程序的加速效率，K 用于衡量程序对数据集规模的适应性。S 与 K 的定义如下：

$$S(m) = \frac{1\text{个节点}1\text{份数据的处理时间}}{m\text{个节点}1\text{份数据的处理时间}} \tag{4}$$

$$K(m) = \frac{1\text{个节点}1\text{份数据的处理时间}}{m\text{个节点}m\text{份数据的处理时间}} \tag{5}$$

对于 MedNIST 数据集，设置式 (4) 与式 (5) 中的 1 份数据为数据集中的 3 000 张医疗图像；对于 Alzheimer 数据集，设置 1 份数据为数据集中的 120 张图像。通过调整 ImageTextInputFormat 参数启动不同数量的 Map 节点，计算 S 与 K 值如图 6 所示。

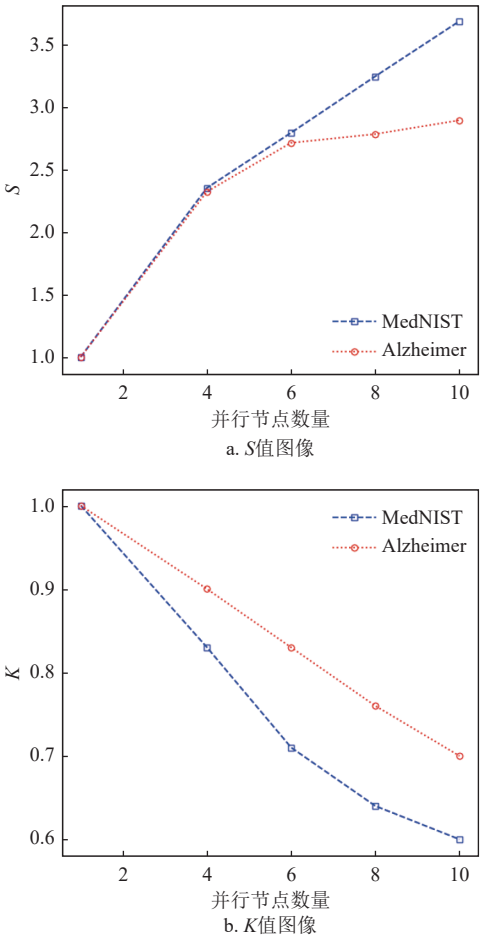


图 6 Map 节点数量不同时的 S 值与 K 值图像

如图 6a 所示，当 Map 节点数量分别为 1、4、6、8、10 时，MIBD2P 程序在 2 个数据集上的加速比 S 随着 Map 节点数量的增长均呈现增长的趋势，表明通过增加并行度可以有效提升程序运行速度。特别地，MIBD2P 的 S 值在 MedNIST 数据集上几乎呈线性增长，表明 MIBD2P 适合处理具有大量图像的数据集；在 Alzheimer 数据集上 S 值的增长速度逐渐放缓，这是因为测试用的 Alzheimer 图像数量相对较少，Map 节点大于 6 时 MIBD2P

程序在该数据集已接近性能极限。观察图 6b，随着 Map 节点和图像数据规模的增加，在 MedNIST 和 Alzheimer 两个数据集上 K 值的下降速度均呈现放缓趋势，MIBD2P 程序对于不同规模的图像数据集具有良好的可扩展性。

在不同的图像数量下，分析 MIBD2P 方案的具体性能。开启固定数量的 Map 节点，改变数据总量，观察程序运行时间；改变 Map 节点数量，对比程序在不同并行度下的运行时间。实验数据如表 3 和表 4 所示。

表 3　MedNIST 数据集上 MIBD2P 程序运行时间

图像	运行时间/s		
总数量/张	6个Map	8个Map	10个Map
240	21.2	21.8	21.0
1 200	31.4	29.2	28.0
6 000	81.6	76.8	66.6
30 000	405.2	338.0	285.8

表 4　Alzheimer 数据集上 MIBD2P 程序运行时间

图像	运行时间/s		
总数量/张	2个Map	4个Map	8个Map
120	37.0	27.6	21.2
240	61.4	39.0	30.4
480	111.8	67.0	45.4

在 Map 节点数量不变的情况下，程序运行时间随着数据量的增加而增加。在图像总数量较少时，程序运行时间并没有与数据量同比例增加，如表 3 中，图像总数量由 240 张增加 4 倍到 1 200 张，程序运行时间并没有同步增长 4 倍。这是因为在图像数量较少时，Mapper 程序实际的计算时间较短，MapReduce 任务配置、各节点的开启关闭占据了大部分的程序执行时间。随着图像数量的增加，如图像数量由 6 000 张增加 4 倍到 30 000 张，MapReduce 任务的实际计算时间逐渐占据总运行时间的主要部分，程序运行时间与数据量开始呈现同比例增加的关系。

观察表 4 的对角线方向，在每个 Map 节点均处理 60 张图像的情况下，程序整体运行时间随着并行运行的 Map 数量的增加而增加，这是因为 Map 数量的增加导致了更多的网络传输和磁盘 IO，最终导致程序整体运行时间增加，与图 6b 中 K 值的变化趋势相符。

观察 MIBD2P 程序在不同并行度时的运行时间，在图像数量较少时，数据处理所需时间较短，任务配置、节点启动关闭的时间开销相对较大，通过增加 Map 节点、提升程序并行度并不能显著改善程序运行效率。随着图像数量的增加，提升程序并行度的效果开始显现，程序运行时间随着 Map 节点数量的增加而大幅减少，与图 6a 中 S 值的变化趋势相符。

5　结束语

本文提出了一种基于多分辨率分析的医疗图像大数据差分隐私保护方案，分析了医疗图像各小波频带不同的隐私保护需求，基于差分隐私指数机制设计扰动算法，根据图像小波频带的隐私需求进行所需强度的差分隐私保护，并在 Hadoop 大数据平台上设计相关算法进行验证。本文通过理论分析证明了所提方案满足ε–差分隐私。在多个医疗图像数据集上进行数据效用和方案性能实验，结果表明本方案在满足差分隐私的同时有效保证了图像质量，相比现有图像差分隐私保护方案能够更好地满足医疗图像对图像质量的要求，可在较大规模的数据集上实现高效的医疗图像大数据差分隐私保护。

未来的工作可以从以下几个方面继续展开：1）对像素差分扰动算法进行优化，提升算法的整体性能；2）在空域内针对医疗图像前景区域和背景区域，根据其隐私保护需求研究相应的差分隐私保护方法；3）除了 SequenceFile 和 TextFile，Hadoop 还支持 RCFile、ORCFile 等文件类型，这些文件类型在数据仓库中的应用更加广泛，后续可以基于 RCFile、ORCFile 等文件类型研究 Hadoop 医疗图像的数据处理，推动医疗图像大数据仓库的建设。

参 考 文 献

[1] MCPHERSON R, SHOKRI R, SHMATIKOV V. Defeating image obfuscation with deep learning[EB/OL]. [2024-04-09]. http://arxiv.org/abs/1609.00408.

[2] FROME A, CHEUNG G, ABDULKADER A, et al. Large-scale privacy protection in google street view[C]//2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009: 2373-2380.

[3] SUN Q, MA L, OH S J, et al. Natural and effective obfuscation by head inpainting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5050-5059.

[4] WANG Z, GUO H, ZHANG Z, et al. Towards compression-resistant privacy-preserving photo sharing on social networks[C]//Proceedings of the 21st International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile

Computing. New York: Association for Computing Machinery, 2020: 81-90.
[5] LIU S, GUO C, SHERIDAN J T. A review of optical image encryption techniques[J]. Optics & Laser Technology, 2014, 57: 327-342.
[6] KAUR M, KUMAR V. A comprehensive review on image encryption techniques[J]. Archives of Computational Methods in Engineering, 2020, 27: 15-43.
[7] VENGADAPURVAJA A M, NISHA G, AARTHY R, et al. An efficient homomorphic medical image encryption algorithm for cloud storage security[J]. Procedia Computer Science, 2017, 115: 643-650.
[8] YANG C N, CHU Y Y. A general (k, n) scalable secret image sharing scheme with the smooth scalability[J]. Journal of Systems and Software, 2011, 84(10): 1726-1733.
[9] LI P, LIU Z, YANG C N. A construction method of (t, k, n)-essential secret image sharing scheme[J]. Signal Processing: Image Communication, 2018, 65: 210-220.
[10] 张旭. 基于差分隐私机制的数据隐私保护的研究[D]. 南京: 南京邮电大学, 2022.
ZHANG X. Research on data privacy protection method based on differential privacy mechanism[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022.
[11] ZHANG X, WANG Y, MA J, et al. QAPP: A Quality-aware and privacy-preserving medical image release scheme[J]. Information Fusion, 2022, 88: 281-295.
[12] DWORK C. Differential privacy[C]//International Colloquium on Automata, Languages and Programming. Heidelberg: Springer, 2006: 1-12.
[13] BEAUDOIN N, BEAUCHEMIN S S. An accurate discrete Fourier transform for image processing[C]//2002 International Conference on Pattern Recognition. Quebec: IEEE, 2002: 935-939.
[14] KHAYAM S A. The discrete cosine transform (DCT): Theory and application[J]. Michigan State University, 2003, 114(1): 31.
[15] DEWITTE S, CORNELIS J. Lossless integer wavelet transform[J]. IEEE Signal Processing Letters, 1997, 4(6): 158-160.
[16] KOR S, TIWARY U. Feature level fusion of multimodal medical images in lifting wavelet transform domain [C]//The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. San Francisco: IEEE, 2004: 1479-1482.
[17] YADAV S P, YADAV S. Image fusion using hybrid methods in multimodality medical images[J]. Medical & Biological Engineering & Computing, 2020, 58: 669-687.
[18] BIBB R, EGGBEER D, PATERSON A. Medical Imaging[M]. Amsterdam: Elsevier, 2015.
[19] FAN L. Image pixelization with differential privacy [C]//Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo: Springer International Publishing, 2018: 148-162.
[20] FAN L. Practical image obfuscation with provable privacy[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai: IEEE, 2019: 784-789.
[21] 张啸剑, 付聪聪, 孟小峰. 结合矩阵分解与差分隐私的人脸图像发布[J]. 中国图像图形学报, 2020, 25(4): 655-668.
ZHANG X J, FU C C, MENG X F. Private facial image publication through matrix decomposition[J]. Journal of Image and Graphics, 2020, 25(4): 655-668.
[22] 张啸剑, 付聪聪, 孟小峰. 面向人脸图像发布的差分隐私保护[J]. 中国图像图形学报, 2018, 23(9): 1305-1315.
ZHANG X J, FU C C, MENG X F. Facial image publication with differential privacy[J]. Journal of Image and Graphics, 2018, 23(9): 1305-1315.
[23] ZHANG G, WEI H, GE L, et al. A differential privacy image publishing method based on wavelet transform[C]//International Conference on Parallel and Distributed Computing: Applications and Technologies. Cham: Springer International Publishing, 2021: 584-595.
[24] CHATZIKOKOLAKIS K, ANDRÉS M E, BORDENABE N E, et al. Broadening the scope of differential privacy using metrics[C]//Privacy Enhancing Technologies: 13th International Symposium. Heidelberg: Springer, 2013: 82-102.
[25] 张良将, 宦飞, 王杨德. Hadoop 云平台下的并行化图像处理实现[J]. 信息安全与通信保密, 2012(10): 59-62.
ZHANG L J, HUAN F, WANG Y D. Parallel image processing implementation under Hadoop cloud platform[J]. Information Security and Communications Privacy, 2012(10): 59-62.
[26] 兰云旭, 王俊峰, 唐鹏. 基于 Spark 的并行医学图像处理研究[J]. 四川大学学报（自然科学版）, 2017, 54(1): 65-70.
LAN Y X, WANG J F, TANG P. Parallel processing researches of medical image based on Spark[J]. Journal of Sichuan University(Natural Science Edition), 2017, 54(1): 65-70.
[27] SWEENEY C, LIU L, ARIETTA S, et al. HIPI: A Hadoop image processing interface for image-based mapreduce tasks[J]. Chris University of Virginia, 2011, 2(1): 1-5.
[28] SHENSA M J. The discrete wavelet transform: Wedding the a trous and Mallat algorithms[J]. IEEE Transactions on Signal Processing, 1992, 40(10): 2464-2482.
[29] SWELDENS W. The lifting scheme: A construction of second generation wavelets[J]. SIAM Journal on Mathematical Analysis, 1998, 29(2): 511-546.
[30] UNSER M, BLU T. Mathematical properties of the JPEG2000 wavelet filters[J]. IEEE Transactions on Image Processing, 2003, 12(9): 1080-1090.
[31] CHRISTOPOULOS C, SKODRAS A, EBRAHIMI T. The JPEG2000 still image coding system: An overview[J]. IEEE Transactions on Consumer Electronics, 2000, 46(4): 1103-1127.
[32] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Berlin: Springer, 2006: 265-284.
[33] MCSHERRY F, TALWAR K. Mechanism design via

differential privacy[C]//Proceedings of the 48th Annual IEEE Symposium on Foundation of Computer Science. Washington: IEEE, 2007: 94-103.

[34] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our data, ourselves: privacy via distributed noise generation[C]//Annual International Conference on the Theory and Applications of Cryptographic Techniques. Berlin: Springer, 2006: 486-503.

[35] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3-4): 211-407.

[36] WHITE T. Hadoop: The definitive guide[M]. Sevastopol: O'Reilly Media Inc, 2012.

[37] SHAPIRO J M. Embedded image coding using zerotrees of wavelet coefficients[J]. IEEE Transactions on Signal Processing, 1993, 41(12): 3445-3462.

[38] WHEELER F W, Pearlman W A. SPIHT image compression without lists[C]//2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul: IEEE, 2000: 2047-2050.

[39] LIAN C J, CHEN K F, CHEN H H, et al. Analysis and architecture design of block-coding engine for EBCOT in JPEG 2000[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(3): 219-230.

[40] 尚涛, 赵铮, 舒王伟, 等. 基于等差隐私预算分配的大数据决策树算法[J]. 工程科学与技术, 2019, 51(2): 130-136.
SHANG T, ZHAO Z, SHU W W, et al. Big data decision tree algorithm based on equal-arrival privacy budget allocation[J]. Advanced Engineering Sciences, 2019, 51(2): 130-136.

[41] PINAYA W H L, TUDOSIU P D, GRAY R, et al. Unsupervised brain anomaly detection and segmentation with transformers[J]. Medical Image Analysis, 2022, 79:102475.

[42] MCAULIFFE M J, LALONDE F M, MCGARRY D, et al. Medical image processing, analysis and visualization in clinical research[C]//Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. Bethesda: IEEE, 2001: 381-386.

[43] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 586-595.

[44] 张晶, 冯林, 王乐, 等. MapReduce 框架下的实时大数据图像分类[J]. 计算机辅助设计与图形学学报, 2014, 26(8): 1263-1271.
ZHANG J, FENG L, WANG L, et al. Real-time big data image classification under MapReduce framework[J]. Journal of Computer-Aided Design & Computer Graphics, 2014, 26(8): 1263-1271.

编　辑　税　红