Collective Protection: Preventing Sensitive Inferences via Integrative Transformation

Dalin Zhang*, Lina Yao*, Kaixuan Chen*, Guodong Long†, Sen Wang‡
*School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
‡Centre for Artificial Intelligence, University of Technology Sydney, Sydney, Australia
‡School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia
{dalin.zhang, lina.yao, kaixuan.chen}@unsw.edu.au, guodong.long@uts.edu.au, sen.wang@uq.edu.au

***Abstract*—Sharing ubiquitous mobile sensor data, especially physiological data, raises potential risks of leaking physical and demographic information that can be inferred from the time series sensor data. Existing sensitive information protection mechanisms that depend on data transformation are effective only on a particular sensitive attribute, together with usually requiring the labels of sensitive information for training. Considering this gap, we propose a novel user sensitive information protection framework without using a sensitive training dataset or being validated on protecting only one specific sensitive information. The presented approach transforms raw sensor data into a new format that has a "style" (sensitive information) of random noise and a "content" (desired information) of the raw sensor data, thus is free of user sensitive information for training and able to collectively protect all sensitive information at once. Our implementation and experiments on two real-world multi-sensor human activity datasets demonstrate that the proposed data transformation technique can achieve the protection for all sensitive information at once without requiring the knowledge of users' personal attributes for training, and simultaneously preserve the usability of the new transformed data with regard to inferring human activities with insignificant performance loss.**

***Index Terms*—mobile sensor, sensitive inference, data transformation, activity recognition**

I. Introduction

Smart wearable devices are key components for human activity recognition (HAR) that encourages various attractive applications of continuous and unobtrusive services in our daily life [1]. However, the prevalence of HAR-based applications is a double-edged sword. On the one hand, it has made our lives easier and much more convenient. On the other hand, it has been raising privacy concerns, i.e., the sensor data from the end users' personal devices are continuously collected and transmitted to the service provider, which might cause severe privacy leakage [2].

Sensitive Information Leakage Example: Consider a scientific study being carried out to investigate the daily activities of users. For this purpose, the users engaging in the study have a smartphone which integrates multiple sensors (e.g. accelerometer and gyroscope) in the pocket to continuously collect data. The collected time series data is shared with the study researchers to infer the users' physical activities, such as walking and running [3]. Due to the difference in personal facts like age, gender and so on, people perform activities in different manners. Thus, these kinds of personal information (age, gender and so on) can also be inferred from the same data that is used for activity recognition [2], [4], [5].

On the one hand, the inference about human activities such as *walking and running* is desired and extremely critical for the study, especially when it is for a healthcare purpose. But, on the other hand, inferences about personal information like *age, gender and so on* are sensitive to users and need to be prevented. Therefore, we draw a conundrum, where the same time series data can be used for making both desired inferences and also sensitive inferences that need to be kept avoided.

Our target in this paper is to propose a data transformation method that prevents all sensitive information from being inferred at once and maintain the desired information being inferred normally. The data transformation module modifies the raw sensor data to get rid of sensitive formation but preserve the desired information, and then the transformed data can be released to a service provider.

Limitations of Current Works: A number of works have studied on preserving data privacy. The differential privacy [6] and its variants [7], [8] provide a strong privacy guarantee by confusing a statistical query response drawn from a population-scale database by adding noise, such that the presence or absence of a user in the database is preserved. Nevertheless, the differential privacy methods are not suitable for protecting sensitive inferences from sensor signals, where an adversary directly interprets sensor signals instead of analyzing statistical query responses.

The existing solutions to preventing sensitive inferences from sensor data are mainly divided into two streams: (1) stopping sensors based on pre-set conditions or user definitions, and (2) transforming raw sensor data to be free of sensitive information before revealed to a service provider. The former solution corrupts sensors to reduce the risk of leaking of sensitive information at the expense of severely decreasing the usability of data [9], [10]. In contrast, the second solution is a more practical and promising approach.

Raw sensor data is first modified by a carefully designed transformation module to eliminate privacy information, and simultaneously preserve the desired information. Then the transformed data is revealed to the service provide for HAR. The data transformation-based solution provides a better privacy-utility balance than the naive sensor corruption ap-

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00197

proach, thus paves a promising way of protecting the sensitive information of sensor data. However, there are two major drawbacks of current data transformation solutions: (1) since current data transformation approaches only specifically target on protecting a certain sort of sensitive information such as gender [11] or identification [2], multiple data transformation models are thus needed to protect multiple sensitive information; (2) the labels of user sensitive information is required for training a data transformation module [2], [11], [12] to directly reduce the sensitive inference accuracy, which is not an ideal solution for practioners.

Our Approach: To tackle the above drawbacks, we propose an integrative framework to protect all user sensitive information at once and the training process is free of the labels of user sensitive information. Inspired by the image style transformation research [13], we regard the user sensitive information like gender or height as the "style" information which influences how a user performs a task, while the desired information which represents what task a user performs is treated as the "content" information. Our framework targets on transforming the data *"style"* comparable to *random noise* but keeping the data *"content"* exact as *raw data*. Specifically, we construct a fully convolutional $TransNet$ for sensitive protection transformation. Different from those methods that directly try to confuse a classifier which is trained for sensitive information inferences using an adversarial training strategy, we design a *style loss* that depends on the features from the $LossNet$ which is pretrained for inferring desired information. The *style loss* tries to minimize the distance between the $LossNet$ features that are extracted from the random noise and transformed sensor data respectively. Meanwhile, to keep the usability of the transformed data, we apply a *content loss*, which minimizes the difference between the $LossNet$ features drawn with the raw sensor data and transformed sensor data. During the training process, only raw sensor data and the labels of desired information are presented while all the labels of privacy-sensitive information are masked. We evaluate the proposed framework on protecting all five kinds of user sensitive information (i.e. age, gender, height, weight, and identification) at once with two multi-sensor human activity recognition datasets. The empirical validation results exhibit that the presented approach reduces the risk of leakage of all sensitive information regarding sensitive inference precision, while remains a high preservation level of desired information with regard to activity recognition accuracy.

II. Methodology

A. Problem Statement and Definition

In this work, we focus on preventing user sensitive information from being inferred from mobile sensor data. As multiple sensing modalities are usually involved in, we first assume all the sensor data is synchronized and recorded at the same frequency. Let $X(t) = [x_1(t), x_2(t), ..., x_m(t)]$ be the reading values of m sensor components (each component could be an axis of a mobile sensor) recorded at time step t. Thus for a time period of d in length starting from time t, we have time-series data $S_d(t) = [X(t); X(t+1); ...; X(t+d-1)]$. For simplicity, we use S_d instead of $S_d(t)$ in the rest of the paper.

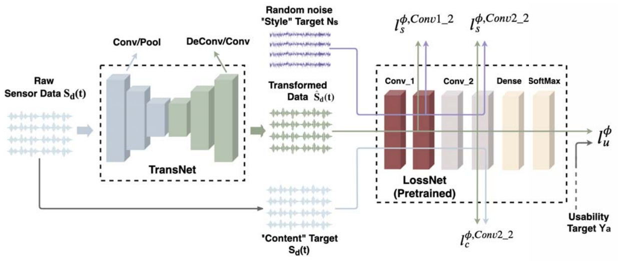


Figure 1. Framework Overview. We first pretrain the LossNet on raw sensor data for inferring desired information. Then the LossNet is fixed, and used to define the loss functions that measure "style" difference between transformed data and random noise and "content" difference between transformed data and raw data. We also define a usability loss to specifically keep the inference accuracy of the desired information. We train the TransNet through minimizing a weighted combination of the above loss functions to protect user sensitive information while simultaneously preserve the desired information.

The S_d is two-dimensional (2D) raw sensor data with one dimension representing time and the other representing sensor components. In traditional conditions, a service provider uses a human activity inference function $I_a(.)$ to recognize users' activities Y_a from raw sensor data S_d for subsequent analysis. Ideally, $I_a(S_d) = Y_a$. At the meantime, there exist an attack function $I_s(.)$ that can be used to infer user sensitive information Y_s (Y_s can be gender, age and so on) from the raw sensor data S_d. Ideally, $I_s(S_d) = Y_s$. Our goal is to find an optimal transformation function $f^*(.)$, so that the sensitive inference cannot be drawn from the optimal transformed data $\hat{S}_d^* = f^*(S_d)$: $I_s(\hat{S}_d^*) \neq I_s(S_d) = Y_s$, whereas the desired inference about human activities $I_n(\hat{S}_d^*)$ can be drawn normally: $I_a(\hat{S}_d^*) = I_a(S_d) = Y_a$. Here $\hat{S}_d$ is the data which is transformed from raw data S_d by a transformation function $f(.)$ and $\hat{S}_d^*$ is the optimal transformed results where the sensitive information cannot be inferred from.

B. Overview

The goal of this study is to transform raw data before being revealed to an untrusted service provider, so that the sensitive information cannot be inferred from the transformed data whereas the desired information can be inferred as usual. To this end, we propose to transform raw data to a new representation that has a "style" (sensitive information) of random noise and a "content" (desired information) of raw data. The transformed data should satisfy such conditions: when an adversary tries to infer user sensitive information from the transformed data, the results should be as unreliable as drawn from random noise; whereas a service provider can make inferences about desired information from the transformed data with as high accuracy as from the raw data.

As shown in Figure 1, in order to achieve such a transformation function $f(.)$, we propose a framework comprising a $TransNet$ $f(.)$ that performs the data transformation process, and a $LossNet$ ϕ that defines several loss functions for training the TransNet. Specifically, the LossNet defines: a "style" loss that measures the "style" difference between the $transformed\ data$ and $random\ noise$, a "content" loss that measures the "content" difference between $transformed\ data$ and $raw\ data$, and a usability loss that specifically helps to keep the inference accuracy of the desired information.

Each loss function computes a scalar value $\ell_i(\hat{S}_d, Y_i)$ measuring the difference between the transformed data $\hat{S}_d$ and a target Y_i (e.g. random noise or raw data). The TransNet is trained with the stochastic gradient descent to minimize the weighted combination of all loss functions:

$$\operatorname{argmin} \mathbf{E}\big[\sum_{i=1} \lambda_i \ell_i(f(S_d), Y_i)\big], \sum_i \lambda_i = 1. \tag{1}$$

C. Network Structure

1) **LossNet:** The LossNet ϕ is a traditional 2D convolutional neural network (CNN) for human activity recognition. The detailed configuration of the LossNet is depicted in Table I. The LossNet has two CNN blocks, each of which has two CNN layers and a maxpooling layer. The input to the LossNet has a size of $m \times d \times 1$, where m is the number of sensing components of raw data, d is time period length, and 1 is the number of feature maps. The period length d is set to 50 in this study. The CNN kernel is set to 1×3 and the maxpooling is always applied along the time dimension to reduce the feature map size by half. After flattening, the output of the second pooling layer is fed into a dense layer of size 400. At last, a dense layer with the softmax activation function defined as softmax(x_i)=$\frac{1}{\mathcal{Z}}$exp(x_i) with $\mathcal{Z}$=$\sum_i$exp(x_i), is appended for final output. The loss function for training the LossNet is a cross-entropy loss for human activity classification:

$$\ell_a^{\phi} = -\sum_e Y_{a,e} log(\phi(S_d)_e), \tag{2}$$

where $Y_{a,e}$ and $\phi(S_d)_e$ is the label and the predicted probability of the activity category e respectively. The predicted probability $\phi(S_d)_e$ is output from the LossNet with the raw data as input.

2) **TransNet:** The TransNet $f(.)$ is a fully convolutional network with downsampling first and upsampling to the original size afterwards. The fully convolutional TransNet can take any size of data as input, which is another advantage of our framework.

The detailed configuration of the TransNet is also illustrated in Table I. We use two convolution/maxpooling pairs to downsample the input raw data followed by two deconvolution/convolution blocks to upsample to the original size. Rather than relying on an interpolating upsampling, deconvolution allows the upsampling process to be learned jointly with the rest of the network. The input and output of the TransNet both have a size of $m \times 50 \times 1$. To achieve the size unchanged, we tune the padding option of both the first pooling layer and the last convolutional layer to the "Valid". The kernel size setting is consistent with the LossNet. The final convolutional layer of TransNet has no activation functions to make the transformed data have possible values both positive and negative.

Table I
THE CONFIGURATIONS OF THE LOSSNET AND TRANSNET. H, W, F, sH, AND sW REFER TO HEIGHT, WIDTH, THE NUMBER OF FEATURE MAPS, STRIDE HEIGHT, AND STRIDE WIDTH RESPECTIVELY.

LossNet				
Layer	Input Size (H×W×F)	Kernel/Stride (H×W/sH×sW)	Padding	Activation
Conv1_1	m×50×1	1×3/1×1	Same	Relu
Conv1_2	m×50×16	1×3/1×1	Same	Relu
MaxPool1	m×50×16	1×2/1×2	Valid	-
Conv2_1	m×25×32	1×3/1×1	Same	Relu
Conv2_2	m×25×32	1×3/1×1	Same	Relu
MaxPool2	m×25×32	1×2/1×2	Valid	-
Dense	flat(m×12×32)	-	-	Relu
Dense	400	-	-	softmax
TransNet				
Conv1	m×50×1	1×3/1×1	Same	Relu
MaxPool1	m×50×16	1×2/1×2	Valid	-
Conv2	m×25×16	1×3/1×1	Same	Relu
MaxPool1	m×25×32	1×2/1×2	Same	-
Conv3	m×13×32	1×3/1×1	Same	Relu
DeConv1	m×13×32	1×3/1×2	Same	Relu
Conv4	m×26×32	1×3/1×1	Same	Relu
DeConv1	m×26×32	1×3/1×2	Same	Relu
Conv4	m×52×32	1×3/1×1	Valid	-

D. "Style" and "Content" Consistency

1) **Content Consistency:** We define a "content" loss function for measuring the "content" consistency between the transformed data and raw data. The "content" information describes what a user does during the data recording period d, which is human activities in this study. We encourage the raw data S_d and the transformed data $\hat{S}_d$ to have similar feature representations as computed by a higher CNN layer of the LossNet ϕ. Formally, let $\phi_j(\hat{S}_d)$ and $\phi_j(S_d)$ be the outputs of the jth layer of the network ϕ when the input of ϕ is the transformed data $\hat{S}_d$ and raw data S_d respectively. $\phi_j(.)$ is of shape $C_j \times H_j \times W_j$. The "content" difference of layer j is defined as the Euclidean distance between the feature representations of the transformed data $\hat{S}_d$ and raw data S_d:

$$\ell_c^{\phi,j} = \frac{1}{C_j H_j W_j} ||\phi_j(\hat{S}_d) - \phi_j(S_d)||_2^2 \tag{3}$$

We use the "content" difference of the layer *Conv2_2* of the LossNet to produce the "content" loss. The "content" loss is:

$$\ell_c^{\phi} = \ell_c^{\phi,Conv2\_2} \tag{4}$$

Using a "content" loss from the intermediate layer of the LossNet to train the TransNet encourages the transformed data to keep the "content" similar to the raw data, but does not force them to match exactly.

2) **Style Consistency:** Besides encouraging similar "content" to raw data, we also would like the transformed data to have a similar "style" to random noise N_s. The "style" represents the manner a user performs a activity which is impacted by personal information like age, gender and so on [14]. These kinds of personal information are sensitive to users and need to be protected. Previous research has reported that a CNN that is originally trained for human activity recognition has the possibility of learning features that could be used for accurately estimating the user personal information, without any intentional design [2]. Thus we here use the LossNet to generate the "style" loss for training the TransNet.

Inspired by the image style transformation process [13], we utilize the Gram matrix to measure the "style" difference. We first give the definition of the Gram matrix. Let $\phi_j(x)$ be the output of the jth convolutional layer of the LossNet ϕ when the input of ϕ is x. The shape of $\phi_j(x)$ is $C_j \times H_j \times W_j$. Then the Gram matrix G_j^{ϕ} can be defined to be a matrix of shape $|C_j| \times |C_j|$ with its elements as:

$$G_j^{\phi}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \phi_j(x)_{h,w,c} \phi_j(x)_{h,w,c'} \quad (5)$$

In practice, the the Gram matrix can be computed easily via $G_j^{\phi}(x) = \Psi\Psi^T / C_j H_j W_j$, where Ψ can be obtained by reshaping ϕ_j into a 2D matrix of shape $C_j \times H_j W_j$. The "style" difference is the squared Frobenius norm of the difference between the Gram matrices of the transformed data $\hat{S}_d$ and the random noise N_s:

$$\ell_s^{\phi,j}(\hat{S}_d, N_s) = ||G_j^{\phi}(\hat{S}_d) - G_j^{\phi}(N_s)||_F^2 \quad (6)$$

The layer Conv1_2 and Conv2_2 of the LossNet are used to produce the "style" loss, which is the sum of the "style" difference of each layer. Therefore, we have the final "style" loss:

$$\ell_s^{\phi} = \ell_s^{\phi,Conv1\_2}(\hat{S}_d, N_s) + \ell_s^{\phi,Conv2\_2}(\hat{S}_d, N_s). \quad (7)$$

3) **Usability Loss.:** We also define a usability loss ℓ_u^{ϕ} to strength maintaining specific desired information during the data transformation process. The usability loss is a cross-entropy loss that measures the difference between the prediction of desired information from the pretrained LossNet with the transformed data as input and the labels of the desired information:

$$\ell_u^{\phi} = -\sum_k Y_{a,k} log(\phi(\hat{S}_d)_k), \quad (8)$$

where $Y_{a,k}$ and $\phi(\hat{S}_d)_k$ is the label and the predicted probability of the activity category k respectively. The activity predicted probability $\phi(\hat{S}_d)_k$ is output from the pretrained LossNet with the transformed data as input.

Thus the final loss function should be the weighted summation of all individual losses λ_c, λ_s^{ϕ}, and λ_u^{ϕ}.

$$\ell^{\phi} = \mathbf{E}\big[\lambda_c \ell_c(\hat{S}_d, S_d) + \lambda_s \ell_s^{\phi}(\hat{S}_d, N_s) + \lambda_u \ell_u^{\phi}(\hat{S}_d, Y_a)\big] \quad (9)$$

The weight of each loss λ_i is set experimentally and kept added up to 1.

E. Training Process

The LossNet ϕ is first trained from scratch on raw training data for inferring desired information that is human activities in this study and then fixed during the subsequent training process of the TransNet. Note that when training the LossNet, only raw training data and the corresponding labels of human activities are provided; the labels of user sensitive information are not required. After training the LossNet, we start to train the TransNet. The goal of training the TransNet is to let the transformed data have a "style" of random noise and a "content" of raw data. Thus the transformed data $\hat{S}_d$, raw data S_d and random noise N_s are input into the pretrained LossNet respectively. Then the final loss calculated by equation 9 is obtained, which at last the TransNet is trained to minimize.

III. Experiment and Results

A. Datasets

We select two public inertial senor-based human activity recognition datasets: MotionSense [11] and MobiAct [15]. Both datasets have five kinds of user sensitive information available: gender (M/F), identification (ID), height (mm), weight (kg), and age (years old). The desired information of both datasets is the activity that a user performs. Thus, specifically the goal of the presented framework is to prevent the sensitive information, namely gender, ID, height, weight, and age from being inferred from the transformed data, while to keep the desired information, namely human activities still being inferred successfully after data transformation.

1) ***MotionSense Dataset:*** The MotionSense dataset has data from two inertial sensors, accelerometer and gyroscope. Four sorts of time series data are obtained from the inertial sensors, namely attitude, rotation rate, user acceleration, and gravity. Each sort of data has three dimensions. Thus there are 12 dimensions in the recording of each time point. A total of 24 users (10 females, 14 males) in a range of gender, age, weight, and height participate in the experiments and collect data of four daily activities: downstairs, upstairs, jogging and walking. Through data inspection, we remove the recordings with data or labels of user information incomplete and finally achieve 264 trials of 767,660 recordings. Following [11], we select 168 long trials of 2 to 3 minutes each for training and the remaining 96 short trials of 0.5 to 1 minutes each for test. After trial segmentation by a 50-length sliding window, we obtain each sample of size (12×50). Finally, there are 61,728 samples for training and 14,098 samples for test.

2) ***MobiAct Dataset:*** The MobiAct dataset comprises data recorded from the accelerometer, gyroscope and orientation sensors of fifty-seven subjects performing nine different types of Activities of Daily Living (ADLs). Different from the MotionSense dataset, there are three kinds of time series data obtained from the sensors, namely orientation, rotation rate, and acceleration (including gravity). Each sort of data also has three axes. Therefore, the recording of each time point has 9 dimensions. After data inspection, we select the data of 44 subjects (14 females, 30 males) performing four ADLs, downstairs, upstairs, walking and jogging, without data and labels of user information missing, in the form of 704 trials of 1,121,296 recordings. As there is no duration difference between the trials of the same activity, we randomly select 66% trials for training and the remaining trials for test. Similarly, we cut each trial with a 50-length sliding window and obtain each sample of size (9×50). Finally, there are 88,412 samples for training and 22,212 samples for test.

B. Experimental Setup

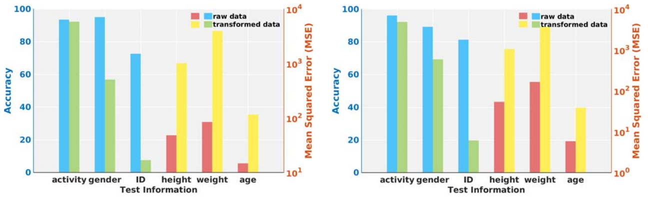
1) **Evaluation Setup:** Following the conventions in previous researches [2], [16], We use the comparison results of inferring both desired and sensitive information before and after transformation to validate the performance of the proposed framework. Specifically, if the accuracy of inferring human activities decreases marginally after data transformation, the proposed framework is regarded as successfully preserving desired information, otherwise is regarded as failing to preserve the desired information. In contrast, if the accuracy of inferring human gender decreases considerably after data transformation, the proposed framework is regarded as successfully protecting the users' gender information, otherwise is regarded as failing to protect the users' gender information. Similar evaluation criteria apply to the other sensitive information.

In order to validate all the sensitive information is protected and the desired information is preserved after data transformation, we build six evaluation neural networks for each dataset for inferring six kinds of information, namely human activity, gender, ID, height, weight, and age. All evaluation neural networks are trained with the same raw training data that is before data transformation and their corresponding ground truth labels. For example, the neural network that is used for evaluating whether the human activity information is unaffected after transformation is trained with the raw training data and human activity labels. Similarly, the evaluation network for gender is trained with raw training data and gender labels.

During the test phase, the raw test data first goes through the well-trained TransNet to achieve the transformed test data, and then the output is fed into each evaluation network respectively to get the evaluation results that are after transformation. For comparison, the raw test data is also fed into the evaluation networks respectively to get the evaluation results of before transformation. The evaluation results of both before and after transformation are presented in the following $Evaluation$ $Results$ section.

With the exception of the activity classifier connected to the dense layer, the architecture of all evaluation networks is the same as that of LossNet. The evaluation networks of activity, gender and ID use softmax output layer for classification. For the numerical information, height, weight, and age, the linear output layer for regression analysis is used. All evaluation networks are trained using the Adam updating rule [17] with a learning rate of 10^{-3}.

2) **Training Setup:** The LossNet is first trained with the trial-independent manner [11] using the raw training data and its paired human activity labels. Afterwards the parameters of the trained LossNet is fixed. Note that the LossNet is only trained with human activity labels, since this information is non-sensitive and can be public. In order to achieve the "style" loss ℓ_s^ϕ for training the TransNet, we generate the noise random noise N_s from random numbers of a uniform distribution between range [-20, 20] to have the same size of the raw data (i.e. 12×50 of MotionSense dataset and 9×50 of MobiAct dataset). The random range is set based on the reasonable scope of sensor readings. The LossNet and TransNet are trained in order using the Adam updating rule [17] with a 10^{-3} learning rate. The weight of each loss function λ_i is experimentally set as $\lambda_s = 0.55$, $\lambda_c = 0.35$, and $\lambda_u = 0.1$.



(a) MotionSense

(b) MobiAct

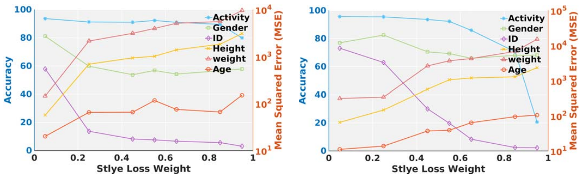
Figure 2. Overall evaluation results on both evaluation datasets. The upper part of this figure displays the absolute results; the bottom part presents the relative results. The activity, gender, and ID are evaluated by accuracy; the height, weight, and age are evaluated by MSE.

C. Evaluation Results

1) **Overall Performance:** Figure 2 plots the overall evaluation results of inferring both desired information and sensitive information with the proposed framework on both evaluation datasets. We use the classification accuracy as the evaluation criteria for inferring categorical information, namely activity, gender and ID, and plot the results with a normal scale to the left axis of Figure 2a and 2b. For the continuous information (i.e. height, weight, and age), we introduce the mean squared error (MSE) as the evaluation criteria and plot the results with a log scale to the right axis of Figure 2a and 2b. As the sensitive information has real-world reasonable ranges (e.g. height, weight and age) and random guess levels (e.g. gender and ID), the absolute results give the intuitive sense about the extend that the proposed framework perturbs the sensitive information.

The results show that our framework obtains satisfactory performance on both datasets with marginal user activity recognition accuracy decrease but significant error increases for inferring all sensitive information. Specifically, after data transformation, the user activity recognition accuracy can still maintain above 90% with only less than 5% drop. This demonstrate that the proposed framework has a nearly-perfect desired information preservation ability. In contrast, when inferring user gender, the accuracy declines dramatically nearly to the random guess level. Note that due to the gender imbalance of the evaluation datasets, the random guess level of gender inference is 58% and 68% for the MotionSense and MobiAct dataset respectively. There is also a considerable drop of ID inference accuracy after data transformation for both datasets. Thus the sensitive information of user gender and ID has been changed to have a random "style" and hard to be precisely inferred after data transformation.

The inference error of the numerical sensitive information, height, weight, and age, also rises remarkably after data transformation. Particularly, inferring user weight experiences the largest performance degradation with MSE increases 46.06 times and 21.36 times after transformation for the MotionSense and MobiAct dataset respectively. Even the smallest performance degradation of inferring user age still suffers more than five times MSE increase. Considering the user age has a relative small reasonable range, the increase of the inference error is significant. The overall results exhibit that our framework learns to transform raw data into a new representation having all sensitive information with a "style" of random noise yet still having the "content" same with that of the raw data.



(a) MotionSense

(b) MobiAct

Figure 3. Privacy-usability tradeoff with different weights of the style loss on both evaluation datasets. The upper part of this figure presents the results of absolute values; the bottom part displays the details of relative results.

2) **Privacy-Utility Tradeoff:** The loss weight λ_i controls the tradeoff between the privacy protection and the usability of transformed data. In this section, we perform experiments by varying the weight of the style loss λ_s^ϕ from 0.05 to 0.95 to investigate the privacy-utility tradeoff of the proposed framework. The content loss and the usability loss change correspondingly and equally share the change value of the style loss to make the overall summation kept to one. Figure 3 shows the results of both the MotionSense and the MobiAct dataset.

It is obvious that with the weight of the style loss growing, the inference error of all sensitive information increases thus the risk of privacy leakage decreases. However, the accuracy of inferring the desired information from the transformed data does not change too much until the style loss weight higher than 0.85. Especially for the MotionSense dataset, the activity recognition accuracy still remains about 90% at 0.85. At the lower style loss weight side, the MSE and inference accuracy of the sensitive information change obviously at 0.25 and fluctuate smoothly afterwards and there is a sharp change after the style loss larger than 0.85. Similarly, the activity recognition accuracy drops clearly at the tail part, that indicates the usability of data decreases significantly at a large style loss weight. The MobiAct dataset has a similar privacy-utility tradeoff trend as the MotionSense dataset with sensitive information protection efficiency goes up apparently after the style weight 0.45. However, the activity recognition accuracy drops a noticeable amount of about 15% with the style weight from 0.65 to 0.85. Finally, the activity inference accuracy sharply downs to only 20% at the end where the transformed data is totally useless.

IV. Conclusion

In this paper, we propose a novel data transformation framework for collectively preventing all user sensitive information from being inferred through mobile sensor data. The presented work transforms raw sensor data into new representations with a "content" same as the raw data but with a "style" like random noise. Different from previous works which can only protect one specific information at once and require the labels of user sensitive information for training, our framework can protect all sensitive information at only one transformation and be trained without requiring the labels of the sensitive information from users. We evaluate the proposed framework on two multi-sensor human activity datasets for protecting all five sorts of user sensitive information. The results demonstrate that our framework is able to protect all user sensitive information at once through the random "style" transformation and to preserve the desired information with a marginal inference accuracy drop.

References

[1] S. Chatterjee, B. Mitra, and S. Chakraborty, "Type2motion: Detecting mobility context from smartphone typing," in *MobiCom*. ACM, 2018, pp. 753–755.

[2] Y. Iwasawa, K. Nakayama, I. E. Yairi, and Y. Matsuo, "Privacy issues regarding the application of dnns to activity-recognition using wearables and its countermeasures by use of adversarial training," in *IJCAI*, 2017, pp. 1930–1936.

[3] K. Chen, L. Yao, X. Wang, D. Zhang, T. Gu, Z. Yu, and Z. Yang, "Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling," in *IJCNN*. IEEE, 2018, pp. 1–8.

[4] J. Lu, G. Wang, and P. Moulin, "Human identity and gender recognition from gait sequences with arbitrary walking directions," *IEEE TIFS*, vol. 9, no. 1, pp. 51–61, 2013.

[5] A. Jain and V. Kanhangad, "Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings," in *ICCTICT*. IEEE, 2016, pp. 597–602.

[6] C. Dwork, "Differential privacy: A survey of results," in *TAMC*. Springer, 2008, pp. 1–19.

[7] F. Liu, "Generalized gaussian mechanism for differential privacy," *IEEE TKDE*, vol. 31, no. 4, pp. 747–756, 2019.

[8] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.

[9] K. R. Raghavan, S. Chakraborty, M. Srivastava, and H. Teague, "Override: A mobile privacy framework for context-driven perturbation and synthesis of sensor data streams," in *PhoneSense*. ACM, 2012, p. 2.

[10] S. Chakraborty, C. Shen, K. R. Raghavan, Y. Shoukry, M. Millar, and M. Srivastava, "ipshield: a framework for enforcing context-aware privacy," in *NSDI*, 2014, pp. 143–156.

[11] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Protecting sensory data against sensitive inferences," in *W-P2DS*. ACM, 2018, p. 2.

[12] S. A. Osia, A. Taheri, A. S. Shamsabadi, M. Katevas, H. Haddadi, and H. R. Rabiee, "Deep private-feature extraction," *IEEE TKDE*, 2018.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711.

[14] T. Brezmes, J.-L. Gorricho, and J. Cotrina, "Activity recognition from accelerometer data on a mobile phone," in *IWANN*. Springer, 2009, pp. 796–799.

[15] G. Vavoulas, C. Chatzaki, T. Malliotakis, M. Pediaditis, and M. Tsiknakis, "The mobiact dataset: Recognition of activities of daily living using smartphones." in *ICT4AgeingWell*, 2016, pp. 143–151.

[16] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Mobile sensor data anonymization," in *IoTDI*. ACM, 2019, pp. 49–58.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.