HUMAN ACTION IMAGE GENERATION WITH DIFFERENTIAL PRIVACY

Mingxuan Sun[*]*, Qing Wang*[*] *and Zicheng Liu*[†]

[*]Louisiana State University, [†] Microsoft Research
msun@csc.lsu.edu, qwang35@lsu.edu, zliu@microsoft.com.

ABSTRACT

Large volumes of human action image data are becoming increasingly available due to the prevalence of surveillance cameras and smart personal devices. While such image data enables important applications such as activity recognition for health and safety enhancement, they often contain sensitive information such as identities that introduce high risks to individual privacy. Existing image privacy-enhancing techniques are either developed at the cost of sacrificing image utility or lack of provable privacy guarantees. We propose a novel human action image generation model that enforces rigorous differential privacy protection. Theoretical analysis is provided to quantify the privacy protection on the training data within the differential privacy framework. Experiments with real-world datasets demonstrate that images generated using our method achieve higher image utilities than baselines given similar degrees of privacy protection.

Index Terms— image synthesis, differential privacy

1. INTRODUCTION

Balancing between human action image utilization and privacy is a fundamental yet challenging problem in computer vision. On one hand, human action images contain richer information than skeleton and depth heatmaps for recognizing activities involving human-object interactions [1]. On the other, those images reveal sensitive information such as gender, race, and identities that can intrude individual's privacy. In this paper, we consider to design a mechanism to transfer raw images to sanitized images so as to assist utilization such as single-image activity recognition [2] while preserving privacy to avoid person identification attacks.

Traditional image privacy-enhancing operations such as blurring, superpixel clustering [3], and downsampling [4] do not consider entangled image factors such as poses and appearances, and thus image utility is largely sacrificed. In human action images, some factors such as pose information are more critical for action recognition and are less intrusive than others such as appearances. Most of those methods are applied in the whole image or in the region of interests such as human faces and bodies to preserve privacy. However, excessive sanitization such as the usage of extreme low resolution (e.g.,16×12) images [5, 6] without considering entangled image factors may reduce data utility.

Generative models have been applied to human action image synthesis [7, 8]. In particular, a framework is proposed in [9] to learn a disentangled representation of the input human body images such as foreground, pose, and background. Novel person images such as images with the same pose but different cloths can be generated by manipulating the new embedding features of each component. However, neither do they provide rigorous privacy guarantees on the training data nor do they generate images with varying degrees of privacy protection.

In this paper, we propose a novel sanitization solution for human action images that provides provable privacy guarantee with minimal impact on utility. Firstly, an image is considered as the composition of several disentangled factors and a sanitization model is learned to generate synthetic images based on less intrusive factors such as skeletons in the raw images. In addition, the model is trained using the original data in a differentially privacy-preserving mechanism, which provides theoretical privacy guarantees for the training data.

In comparison with current approaches, our proposed solution has the following contributions: (1) Based on controllable differential privacy parameter ϵ, the framework can generate human action images conditioned on poses with varying degrees of privacy protection. (2) The privacy cost is analyzed within the framework of differential privacy without assuming any prior knowledge on specific privacy tasks. (3) The results show empirically that our method performs better than baselines in terms of balancing between utility and privacy.

2. RELATED WORK

Recent learning based approaches specify utility tasks (e.g., action recognition) and privacy tasks (e.g., gender inference), define the associated privacy costs (e.g., gender classification accuracy), and formulate the utility-privacy tradeoff as a min-max optimization problem [10, 4]. In particular, a function is learned to transfer raw input (e.g., images, videos) to a sanitized version so as to preserve performance on the utility task and lower performance on the privacy task. For example, in [4], the utility task is activity recognition and the privacy tasks are the classifications of private attributes such as gender, age,

978-1-7281-1331-9/20/$31.00

and hair color. In [10], the utility task is face expression classification and the privacy task is face identity inference. However, those task-driven methods need prior knowledge such as specific privacy attackers and training labels, which may be unknown ahead of data releasing.

Differential privacy [11] has been the gold standard for sanitizing and releasing statistical datasets. It ensures that an adversary is less likely to distinguish between two datasets/databases that differ in at most one element, by observing the output of certain private algorithms. Recent studies [12, 13, 14] have proposed to sanitize image data with such rigorous privacy guarantees. For example, Dp-GAN [13] is a deep generative model trained using the original data enforcing differential privacy principles. However, most of those privacy preserving techniques assume an image as a set of pixels, and thus the protection mechanism does not take advantages of the disentangled image factors such as appearances and poses.

3. BACKGROUND AND PRELIMINARIES

Disentangled image generation: frameworks such as [9] have been proposed to synthesis novel images by manipulating new embedding features of independent factors including appearance and pose. The framework contains two stages. In stage I, given a person's image and the pose keypoint heatmap, the image is divided into three main factors such as foreground, background and pose. A network is constructed to encode each factor to a latent feature, which are then combined through a decoder to re-construct the input image. In stage II, for each factor, a mapping function is trained in an adversarial manner to map Gaussian noises to the latent feature space learned in Stage I.

Differential privacy: a mechanism to provide theoretical guarantees against to what extent an adversary can distinguish adjacent datasets by observing the results of some randomized algorithm. In our cases, each training dataset consists of a set of images. Two of these datasets are considered as adjacent if only one image is present in one dataset but not in the other.

Definition: let $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ be a randomized algorithm that maps domain $\mathcal{D}$ to range $\mathcal{R}$. Let $d, d' \in \mathcal{D}$ be two adjacent datasets that differ in at most one entry. The algorithm $\mathcal{M}$ is said to satisfy (ϵ, δ)-differential privacy, where $\epsilon \geq 0$ and $\delta \geq 0$, if for any two adjacent datasets $d, d' \in \mathcal{D}$, and for any subset $R \subseteq \mathcal{R}$, the following holds true:

$$\Pr[\mathcal{M}(d) \in R] \leq e^{\varepsilon} \Pr\left[\mathcal{M}\left(d'\right) \in R\right] + \delta. \qquad (1)$$

The above definition is a relaxed variant of $(\epsilon, 0)$-DP [11].

Given a deterministic real-valued function $f : \mathcal{D} \rightarrow \mathcal{R}$, a common approach to achieve differential privacy is to add addictive noise to the output according to f's sensitivity Δf. The sensitivity is defined as the maximum difference of the outputs of two adjacent datasets d and d', i.e., $\Delta f = \max_{d \sim d'} |f(d) - f\left(d'\right)|$. Gaussian mechanism, which is a common choice to perturb the output, is defined as:

$$\mathcal{M}(d) = f(d) + \mathcal{N}\left(0, (\Delta f)^2 \sigma^2\right), \qquad (2)$$

where $\mathcal{N}\left(0, (\Delta f)^2 \sigma^2\right)$ is the normal distribution with zero mean and standard deviation $\Delta f \sigma$.

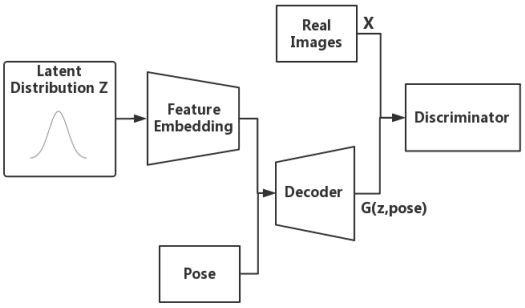

Fig. 1. Our framework consists of a two-step generator and a discriminator. The generator first maps Gaussian variables to a latent feature space and then maps latent features to images conditioned on poses. The generator has no direct access to real images except pose keypoint heatmaps. Compared to existing frameworks, ours is simpler and is capable of rigorous differential privacy analysis.

4. MODEL AND ALGORITHM

We introduce our novel method that integrates human action image generator with differential privacy mechanism.

4.1. Privacy-Preserving Generator

As shown in Figure 1, our generator consists of two-step mapping functions to map Gaussian noises to the image space. Firstly, a feature mapping function $\phi(z)$ is learned to map a Gaussian space to a latent feature embedding space $e = \phi(z)$. Secondly, a decoder is learned to map the feature embedding space to the real image space $\hat{x}$ conditioned on pose keypoints (i.e., skeleton coordinates). In the end, a discriminator tries to differentiate generated images from real images. After training, the generator can synthesize a random image conditioned on the pose of a source image. In this way, the generated image preserves skeleton/pose information as much as possible while obfuscating other factors such as appearances.

Compared to existing disentangled frameworks such as [9], ours is simpler and is capable of supporting differential privacy analysis. Firstly, the generator loss in existing framework includes an image reconstruction error, which means the generator has direct access to original training images. However, the differential privacy costs due to data access from a complicated generator is hard to track. In our framework, the generator has no direct access to real images except pose keypoint heatmaps. Secondly, existing framework requires preprocessing such as decomposing images to multiple factors

like body segments using masks. However, steps involving extra image segmentation on original images introduce extra privacy costs and the cost is hard to quantify theoretically. In our framework, there are no pre-processing stages.

Specifically, in the first step, the intuition of mapping a noise to a feature embedding space instead of an image space is as follows. It has been shown that images usually lie in a low-dimensional space (i.e., feature space). The distribution of the feature space is more continuous and easier to learn compared to the original high-dimensional image space.

In the second step, conditioned on pose keypoint heapmaps, a decoder maps feature embeddings to real images. Specifically, we adopt a person pose estimator [9] to obtain pose heatmaps from original images. A pose keypoint heatmap is a 18-channel image where each channel is an intensity map for one keypoint coordinate. Further, the heatmap is concatenated with feature embeddings and passed into a convolutional autoencoder with skip connections, namely "U-Net"-based architecture, to generate the final image. The combination of feature embeddings and pose enforces the network to learn to synthesize appearance for each pixel with the guidance of the pose keypoints.

The network structure is described in Table 1. The feature mapping is implemented with 4 residual blocks. The decoder is implemented as a U-Net structure with 8 residual blocks. Each residual block's structure is shown in Table 1. Our discriminator is composed of four convolution layers with filter size increasing from 64 to 512 by doubling filter number every time. Leak rectified linear units (LeakyReLU) with $\alpha = 0.2$ is applied after each convolution layer. Dropout layers are used after the activation function, with 0.25 dropout rate.

4.2. Optimization

In order to optimize the proposed model, we apply Wasserstein GAN loss to minimize the earth mover distance between the generated image distribution and the real image distribution. Compared to the original GAN, Wasserstein GAN improves training stability and convergence rate. Formally, the loss functions for Generator G and Discriminator D are:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{\text{data}}(x)}[D(x)] - \mathbb{E}_{z\sim p_z(z),p\sim p_{\text{pose}}(p)}\left[D\left(G(z,p)\right)\right], \quad (3)$$

$$\mathcal{L}_G = \mathbb{E}_{z\sim p_z(z),p\sim p_{\text{pose}}(p)}\left[D\left(G(z,p)\right)\right], \quad (4)$$

where z is the Gaussian noise and p is the pose keypoint heatmap.

Since our human action image generator is more complicated than a vanilla GAN, it is expected that the training process may converge slower. More training iterations mean more privacy cost since each data access introduces additional cost. To mitigate this problem, we propose to leverage pre-trained model on public data to initialize our model parameters in a way similar to the warm start trick introduced in [13]. Specifically, we adopt the decoder in Stage I and the pre-trained feature mapping function in Stage II of the model described in [9]. We will continue the training of our model under the differential privacy constraint and track the privacy cost. The pre-training strategy makes the model easier to converge and also saves certain amount of privacy budget. Note that the existence of public data such as research lab human action data is useful but fairly limited. It is thus critical to adopt our privacy-preserving model for a large amount of private data such as sensor images in homes and hospitals.

Table 1. Network architecture. "×2" means repeating the structure twice. f: filter size, k: kernel size, s: stride length. n_1 starts from 128 and increases by 128 each time until it reaches 640. n_2 starts from 512 and decrease by 128 every time until it becomes 128.

Feature Mapping	
Layer	
1	512 Fully-Connected, ReLU
2-5	(512 FC, ReLU)×2
6	128 FC
Decoder	
1-5	(Conv2D- n_1f-3k-1s, ReLU)×3
6	(Conv2D-640f-3k-1s, ReLU)×2
7-10	Upsampling-2s, Conv2D-n_2f-1k-1s, ReLU (Conv2D-n_2f-3k-1s, ReLU)×2
11	Conv2D-3f-3k-1s, Tanh
Discriminator	
1-4	Conv2D-n_3f-3k-2s, LeakyReLU, Dropout
5	1 Fully-Connected

4.3. Differentially Private Training

Inspired by the previous work [12, 13], we utilize differential privacy to enhance the privacy protection in our proposed model by injecting random noises in the optimization procedure. Specifically, random noise sampled from Gaussian distribution is added to the gradients of discriminator in regard to training images. There are two reasons that we do not add perturbations to the two-step generator, i.e., feature mapping and decoder. First, only the discriminator has access to real images, and thus perturbations in training the discriminator is sufficient for controlling the privacy. Even though the generator has access to pose keypoint heatmaps, those are not intrusive and hence not considered in the privacy protection. Second, compared to generators which may have batch normalization and residual layers, the discriminators are relatively simple and the privacy cost can be tightly estimated. Based on the theorems of differential privacy, we can estimate the privacy cost each time we have access to the training data, and the cumulative privacy loss in the training process.

The construction of our deferential privacy preserving model is outlined in Algorithm 1. In each step, a batch of data are randomly sampled from the original dataset. Specifically, lines from 2 to 3 describe the two-stage generator, which first

Algorithm 1: Differential Private Disentangled GAN

Input: batch size m, training size M, learning rates λ_g, λ_d, iterations n_d, clipping parameter c, gradient norm bound C, noise scale σ, total privacy budget (ϵ, δ), pre-trained feature mapping and decoder if available.

Output: Differential private generator G

while θ *not converge* **do**

  for $l = 1, ..., n_d$ **do**

    for $i = 1, ..., m$ **do**

      1. sample $x \sim p_{real}, y \sim p_{pose}, z \sim p_z$;

      2. $e = \phi(z)$;

      3. $\hat{x} = Dec(e, y)$;

      // compute gradients of discriminator

      4. $g_w^i \leftarrow \nabla_w \left[f_w\left(x^{(i)}\right) - f_w\left(g_\theta\left(z^{(i)}, y\right)\right)\right]$;

      // clip gradients

      5. $g_w^i = g_w^i / max(1, \frac{||g_w^i||_2}{C})$

    // perturbation

    6. $\overline{g}_w \leftarrow \frac{1}{m}\left(\sum_{j=1}^{m} g_w\left(\mathbf{x}^{(j)}, \mathbf{z}^{(j)}\right) + N\left(0, \sigma^2 C^2 I\right)\right)$;

    7. compute cumulative privacy loss according to moments accountant.

    8. $w \leftarrow SGD(w, g_w)$;

  9. sample $y \sim p_{pose}$,$z \sim p_z$;

  // update generator parameters

  10. $g_\theta \leftarrow -\nabla_\theta \frac{1}{m} \sum_{j=1}^{m} f_w\left(g_\theta\left(z^{(j)}, y\right)\right)$;

  11. $\theta \leftarrow \mathrm{SGD}\left(\theta, g_\theta\right)$;

  12. update $\hat{\delta}$ according to ϵ, if $\hat{\delta} \geq \delta$, break;

maps a Gaussian noise to a feature embedding and then to an image based on the pose. At line 4, we calculate the gradients of discriminator with respect to a random subset of images. After that we clip the gradient of the discriminator by a threshold C (line 5) and then perturb the gradient with a Gaussian noise (line 6). In addition, we employ moments accounting [12] to track the privacy budget (ϵ, δ), which is accumulated every time we inject noise to gradients. The parameters of discriminator and generator are dynamically updated until convergence or reaching the privacy budget.

According to standard arguments [11], if we choose the Gaussian noise σ in Algorithm 1 to be $\sqrt{2 \log \frac{1.25}{\delta}}$, the procedure after each batch achieves (ϵ, δ)-DP with respect to the sampled data in the batch. The random sampling of each batch provides additional level of privacy protection. According to the privacy amplification theorem [12], the subsampling procedure achieves $(q\epsilon, q\delta)$-DP with respect to the entire dataset, where $q = m/n$ is the sampling ratio per batch and $\epsilon <= 1$. Further, the accumulated privacy costs after all iterations can be estimated based on moments accounting.

Theorem 1. *Given the sampling ratio $q = m/n$ and the number of total iterations T, there exist constants c_1 and c_2 so that Algorithm 1 achieves (ϵ, δ)-DP for any $\epsilon < c_1 q^2 T$ and $\delta > 0$ if the noise scale σ and the clipping threshold C are chosen appropriately.*

Proof. Let f denote the gradient update function which maps data samples to a real valued vector and the output is bounded by a constant C, e.g., $||f||_2 \leq 1$. Let M be the random Gaussian mechanism $M(d) = \sum_{i \in S} f(d_i) + \mathcal{N}(0, \sigma^2 \mathbf{I})$ where S represents a subset data in a batch.

Let L denote the privacy loss, which is a random variable and is defined as:

$$L\left(o; \mathcal{M}, d, d'\right) = \log \frac{\Pr[\mathcal{M}(d) = o]}{\Pr\left[\mathcal{M}\left(d'\right) = o\right]}, \tag{5}$$

where d, $d' \in \mathcal{D}^n$ are two adjacent datasets, and $o \in \mathcal{R}$ is an output. The privacy loss can be estimated by the λ^{th} moment of L, which is defined as follows:

$$\alpha_{\mathcal{M}}\left(\lambda; d, d'\right) = \log \mathbb{E}_{o \sim M(d)}\left[\exp\left(\lambda L\left(o; \mathcal{M}, d, d'\right)\right)\right]. \tag{6}$$

Further, we need to bound all possible values of $\alpha_{\mathcal{M}}$ and we define $\alpha_{\mathcal{M}} \triangleq \max_{d,d'} \alpha_{\mathcal{M}}\left(\lambda; d, d'\right)$ as the maximum value over all possible adjacent datasets d, d'. It can be proven that for a Gaussian mechanism M, $\alpha_{\mathcal{M}}$ is bounded by:

$$\alpha_{\mathcal{M}}(\lambda) \leq q^2 \lambda(\lambda + 1)/(1 - q)\sigma^2 + O\left(q^3/\sigma^3\right). \tag{7}$$

In addition, $\alpha_{\mathcal{M}}$ has two other properties [12]: composability and tail bound. Composability indicates that if a mechanism $\mathcal{M}$ is composed of a series of sub-mechanisms $\mathcal{M}_1, ..., \mathcal{M}_k$, we have $\alpha_{\mathcal{M}}(\lambda) \leq \sum_{k=1}^{K} \alpha_{\mathcal{M}_k}(\lambda)$. Tail bound indicates that the mechanism $\mathcal{M}$ meets (ϵ, δ)-dp if $\delta = \min_\lambda\left(\alpha_{\mathcal{M}} - \lambda\epsilon\right)$ with $\epsilon > 0$.

According to the two properties above and (7), the log moment of Algorithm 1 can be bounded by $\alpha(\lambda) \leq q^2 \lambda^2 t / \sigma^2$. Therefore, Algorithm 1 is (ϵ, δ)-differential private as long as the following conditions are satisfied: (1) $Tq^2\lambda^2/\sigma^2 \leq \lambda\varepsilon/2$, (2) $\exp(-\lambda\varepsilon/2) \leq \delta$, and (3) $\lambda \leq \sigma^2 \log(1/q\sigma)$. It can be proven that there exist some constants c_1 and c_2 such that when we choose σ to be $\sigma \geq c_2 q\sqrt{T \log(1/\delta)}/\varepsilon$ for any $\varepsilon \leq c_1 q^2 T$, the above three conditions hold. □

5. EXPERIMENT AND RESULTS

5.1. Dataset and Baselines

3We use CAD-60 [15] and Market-1501 [16]. CAD-60 is a human activity dataset and consists of four different persons performing twelve activities in the indoor environment, resulting in 60 RGB videos. We choose four activities, i.e., talking on the phone, drinking water, opening pill container, and writing on whiteboard. The total number of image frames from videos we use is 18960. Market-1501 has 32668 images of 1501 persons which includes 12936 training images of 751

Fig. 2. Compare original images, skeleton images, and sanitized images using different methods on CAD-60 (top 2 rows) and Market-1501 (bottom 2 rows).

persons and 19732 testing images of 750 individuals. We use all 12936 training images for image generation.

Baselines include blurring and superpixel in [3] as well as image downsampling in [4]. All methods have parameters to control image obfuscation levels. We choose 5 different parameter values for each method. For blurring, we use Gaussian blur with five kernel size, i.e., 1, 3.8, 4.4, 10 and 15. The bigger, the more obfuscated. For downsampling, the scale is set to 5, 8, 11, 14 and 50. The larger the scale, the blurrier. For image superpixeling, the segmentation number is set to 2000, 800, 20, 14 and 1. The smaller, the more obfuscated. For our method, we use the same set of poses as conditions and train generation models with different privacy parameter ϵ e.g., 10^6, 20, 5, 1, 0.8. As ϵ decrease, more Gaussian noises will be injected to the model and thus the generated images have more privacy protection.

When training our model, we set other parameters as follows: $\delta = 10^{-5}$, batch size $m = 10$, generator learning rate $\lambda_g = 8e^{-7}$, discriminator learning rate $\lambda_d = 8e^{-7}$, the number of discriminator iterations per generator iteration $n_d = 5$, clipping parameter $c = 0.01$, and gradient bound $C = 1$.

5.2. Balance between Utility and Privacy

We demonstrate that our method can achieve higher privacy protection with minimal impact on utility. For the utility task, we choose single-image action recognition because it is a core computer vision task with applications such as smart homes. The quantification of privacy may vary by contexts, and we choose a privacy protection task to avoid the attack of person identities similar to [4, 10]. We emphasize that the proposed generation framework does not assume specific target utility or privacy tasks.

Table 2. Compare identity attack accuracy on images protected with different methods that achieve utility accuracy around 65%.

Datasets	Methods			
	Our $\epsilon = 1$	Blur=3.8	Down=8	Pixel=800
CAD-60	**0.2862**	0.8020	0.8802	0.9731
Market-1501	**0.0027**	0.1305	0.3901	0.6818

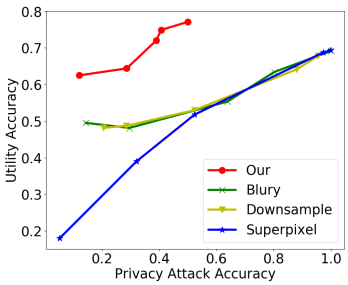

Fig. 3. Compare privacy-utility trade-off on CAD-60 dataset.

For the utility evaluation, each video of CAD-60 dataset has a corresponding activity label and the label is applied for each image frame. We then apply each of the 4 image obfuscating methods with one of 5 parameters to raw images and get a set of obfuscated image sets. For each set, we use the images of three persons as training and the images of the fourth person as testing. We utilize 4-fold cross-validation to evaluate the classification performance of a single image recognition classifier, e.g., a baseline classifier in [2].

For the privacy evaluation, we consider a person identification attack in a simple scenario [10] where the attacker has access to the original training images with the corresponding person identities. However, the attacker cannot see original test image but only the privacy-protected version. In such scenario, the attacker can utilize the training images to train an identification classifier. During testing stage, we use different protecting methods to sanitize the original testing images, resulting in several sanitized testing datasets. The trained classifier is then used to identify these sanitized testing images.

To compare different methods, we show images obfuscated by each method given certain parameter values that achieve the same utility accuracy. For example, in order to achieve 65% in activity recognition, the kernel size of image blurring needs to be 3.8, the scale of downsampling needs to be 8, the segmentation number of image superpixel is 800, and ϵ of our method is 1. Note that it is not suitable to perform action classification on Market-1501 dataset. For consistency, we adopt the same parameters as used in CAD-60. Samples of the obfuscated images under these settings are shown in

Figure 2. We can see that our method is more effective in preserving privacy (e.g., more gender/appearance ambiguity) when achieving the same level of utility.

Table 2 summarizes the privacy attack accuracy achieved by different protecting methods. In particular, with parameters that achieve the action recognition accuracy at around 65%, our proposed model achieves person identification accuracy at as low as 0.2862 for CAD-60 while the accuracy of baselines all exceed 0.8. As for Market-1501, the person identification accuracy of our method is 0.0027, which is close to the random guess 0.0013 (since there are 751 persons) while others are much higher. Our proposed method effectively protects identities compared with baselines.

When image protection level decreases (e.g., less blurred images), the utility accuracy will increase. Meanwhile, the images are less robust to privacy attacks thus attack accuracy will also increase. Thus, the utility accuracy should be a monotonically increasing function with respect to identity attack accuracy. Figure 3 shows the performance of our method and baselines on the CAD-60 dataset with different parameter values. The x-y coordinates of each marker indicate the attack accuracy and utility accuracy of images with the choice of one parameter value. For example, as our privacy protection level decreases (e.g., $\epsilon = 0.8$ to $\epsilon = 10^6$), both the action utility and identity attack accuracy increase. We can observe that at the same level of identification attack, our method always achieves higher utility than baselines.

6. CONCLUSIONS

We proposed a novel sanitization framework that is able to generate synthetic human action images with provable privacy guarantees. Experiments demonstrate that our method achieves high utility in tasks such as single-image activity recognition under similar level of privacy protection. Our framework can be applied to those video synthesis tasks where we can extract each individual frame from the video and sanitize images independently. In the future work, we would like to extend our technique to handle temporal smoothing constraints with rigorous privacy guarantees.

7. ACKNOWLEDGEMENT

This work was supported in part by the National Science Foundation under Grant No.1927513 and the Louisiana Board of Regent under Grant No. LEQSF (2017-20)-RD-A-29.

8. REFERENCES

[1] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 1290–1297.

[2] Zhichen Zhao, Huimin Ma, and Shaodi You, "Single image action recognition using semantic body part actions," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3391–3399.

[3] Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 27–34.

[4] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin, "Towards privacy-preserving visual recognition via adversarial training: A pilot study," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 606–624.

[5] Ji Dai, Behrouz Saghafi, Jonathan Wu, Janusz Konrad, and Prakash Ishwar, "Towards privacy-preserving recognition of human activities," in *2015 IEEE international conference on image processing (ICIP)*, 2015, pp. 4238–4242.

[6] Michael S Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2017.

[7] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang, "Skeleton-aided articulated motion generation," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 199–207.

[8] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool, "Pose guided person image generation," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 406–416.

[9] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz, "Disentangled person image generation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 99–108.

[10] Jiawei Chen, Janusz Konrad, and Prakash Ishwar, "Vgan-based image representation learning for privacy-preserving facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1570–1579.

[11] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2006, pp. 486–503.

[12] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[13] Xinyang Zhang, Shouling Ji, and Ting Wang, "Differentially private releasing via deep generative model (technical report)," *arXiv preprint arXiv:1801.01594*, 2018.

[14] Liyue Fan, "Practical image obfuscation with provable privacy," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 784–789.

[15] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena, "Human activity detection from rgbd images," in *Workshops at the twenty-fifth AAAI conference on artificial intelligence*, 2011.

[16] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.