Towards A Framework for Privacy-Preserving Pedestrian Analysis

Anil Kunchala[1], Mélanie Bouroche[2] and Bianca Schoen-Phelan[3]
Technological University Dublin[1,3], Trinity College Dublin[2], Ireland
d20125529@mytudublin.ie[1], melanie.bouroche@tcd.ie[2], bianca.schoenphelan@tudublin.ie[3]

Abstract

The design of pedestrian-friendly infrastructures plays a crucial role in creating sustainable transportation in urban environments. Analyzing pedestrian behaviour in response to existing infrastructure is pivotal to planning, maintaining, and creating more pedestrian-friendly facilities. Many approaches have been proposed to extract such behaviour by applying deep learning models to video data. Video data, however, includes an broad spectrum of privacy-sensitive information about individuals, such as their location at a given time or who they are with. Most of the existing models use privacy-invasive methodologies to track, detect, and analyse individual or group pedestrian behaviour patterns. As a step towards privacy-preserving pedestrian analysis, this paper introduces a framework to anonymize all pedestrians before analyzing their behaviors. The proposed framework leverages recent developments in 3D wireframe reconstruction and digital in-painting to represent pedestrians with quantitative wireframes by removing their images while preserving pose, shape, and background scene context. To evaluate the proposed framework, a generic metric is introduced for each of privacy and utility. Experimental evaluation on widely-used datasets shows that the proposed framework outperforms traditional and state-of-the-art image filtering approaches by generating best privacy utility trade-off.

1. Introduction

Pedestrians and pedestrian-friendly infrastructure plays a crucial role in planning and creating sustainable urban transportation. Analyzing pedestrian behavior in response to existing infrastructure is pivotal to planning, maintaining, and creating more pedestrian-friendly environments. For example, by analyzing long-term patterns of pedestrian movement and behavioural patterns, local government and business owners can also gain insights for redesigning their space to better manage and plan more pedestrian-friendly facilities.

Visual data is typically used to perform pedestrian analysis in artificial intelligence technologies. Having access to large amounts of open data has facilitated the rapid development of these technologies. As an example, datasets such as ImageNet [12], which contains approximately 14 million labeled images and can be used to train image classification algorithms, have had a significant impact on a wide range of research applications, ranging from self-driving cars to computers 'learning' to play video games, among many others. While the widespread availability of open data benefits research, public datasets have recently been called out for scrutiny [47, 56] due to serious privacy concerns[60, 49] and the potential for data misuse [56, 18]. These concerns are prevalent in pedestrian datasets [53, 12, 55, 2, 9], where the data typically comprises surveillance videos. As a result, some datasets are no longer available to researchers [53, 2], while others are only available under very restricted licenses [55]. This severely limits the amount of data that can be used to perform innovative research.

To address privacy concerns associated with datasets, the research community has developed a number of techniques for protecting individuals' privacy information prior to publishing datasets. The majority of the current research on visual privacy-preserving algorithms focuses on traditional image filtering techniques (such as blurring and pixelation) to obfuscate and/or degrade the sensitive regions of individuals (like their faces). There is a growing body of evidence, however, that suggests that a wealth of information can be mined and a person can still be identified even if their face is blurred, pixelated or completely covered with a blackbox [8]. Conversely, obfuscating all the private information in a dataset will affect its utility. Because of the negative correlation between privacy and utility, it is critical to find ways to maximize datasets utility while minimizing privacy concerns. Recently, visual abstraction methods based on segmentation [40] and neural-art [8] have been developed to improve privacy while maintaining utility, by replacing people with their silhouettes or neural-art based style images. These methods, however, are limited to a single pedestrian per image and do not investigate the impact of privacy-preserving algorithms on the pedestrian's pose, shape, or image background information. Moreover, privacy preser-

vation is evaluated using perceptual studies, which are subjective and do not evaluate the machine understanding of privacy. Finally, utility calculation is only limited to pedestrian count [7], which does not represent a dataset's generalized utility accurately, as most applications, such as activity recognition and behavior prediction, require keypoint information in addition to the the number of people in the image.

Motivated by these limitations, this paper introduces an end-to-end framework by integrating state-of-the-art 3D body and shape estimation with digital in-painting to provide privacy-enhanced representations that capture each pedestrian's presence, pose, shape, and visual background information. In contrast to the perceptual studies, a statistical similarity measure is introduced to quantify privacy using a Siamese Convolutional Neural Network (Si-Net). Inspired by FaceNet [57], a statistical similarity of image embeddings between original and privacy enhanced images is used to quantify machine understanding of privacy-enhanced data. The proposed framework is evaluated on challenging single-person and multi-person per frame pedestrian datasets to demonstrate that it achieves an improved privacy utility trade-off.

The contributions of this paper are as follows:

1. A novel end-to-end framework is introduced to generate a privacy-enhanced version of a given video or image sequence.

2. Both a generic utility and statistical similarity-based privacy metrics are proposed to evaluate the privacy utility trade-off.

We evaluate the proposed framework using the PeVID [28], CMU [1], MOT_16 [34] and i.c.sens [38] datasets, and show that our proposed privacy preserving framework outperforms state-of-the-art baseline methods by achieving the best privacy, while preserving utility.

The remainder of the paper is organised as follows: Section 2 explores the related work in visual privacy-preserving methods, wireframe representation of human bodies, and video in-painting. Section 3 describes the proposed privacy-preserving framework, including wireframe generation and background extraction. Section 4 introduces the privacy and utility metrics including the Si-Net. Section 5 presents the evaluation, followed by Results in Section 6. Finally, Section 7 concludes the proposed work and delineates potential future work.

2. Related Work

This section discusses the existing literature focusing first on visual privacy-preserving methods, followed by privacy & utility metrics, then human body wireframe representation, and eventually video in-painting technologies.

2.1. Visual Privacy-Preserving Methods

Visual privacy-preserving methods hide some of the original information in line with regulatory data protection laws and individual privacy needs. Image filtering techniques such as blurring [15, 36, 66, 35, 39] and pixelation [6, 25, 36] are widely used to improve privacy. Image filters can be applied to entire images or to specific areas of an image that require privacy. The majority of research in visual privacy protection is based on methods that manipulate and/or remove information from faces in images [24, 39, 8, 13]. Personal data information is not limited to faces [8], however, and therefore face obfuscation alone does not guarantee privacy [63]. The person can be easily identified using visual ques, such as clothing information even if face is completely covered with black-box. Visual data should obfuscate all private information to provide complete privacy. In practice, removing all information is not possible, and the dataset might become unusable as a result [63]. Recently, neural-art based obfuscation [7] has been proposed, in which a person's style is altered to protect their identity while preserving naturalness and the data's utility. Visual abstraction methods [11, 24, 8, 7] have recently gained popularity as a way to maintain dataset utility while improving privacy. The goal of visual abstraction methods is to protect people's privacy by replacing their entire body with a generated abstract visual model. Commonly used visual abstraction models include silhouettes [40, 42, 11], bounding box [43] , and avatars [10, 5, 8, 7, 24]. One particularly interesting visual abstraction method is 3D wireframe representation [4]. Although they are similar to avatar representation, wireframes are used as a complete abstraction of the user's identity while preserving the body shape and pose. Despite the fact that wireframes provide a detailed representation of the human body [54, 48, 27], there is a research gap regarding the effectiveness of privacy-preserving and utility capabilities.

2.2. Existing Privacy & Utility Metrics

Privacy and utility can be evaluated using subjective and objective evaluation methods. Through user perception studies and questionnaires, subjective evaluation is used to determine the privacy or utility of a video [8, 40, 13]. Subjective evaluation methods are quite common, and the results of user perception studies may depend on the study group and their assessment of the quality. Subjective evaluations are constrained. For example, evaluating all videos or images in a large public dataset (such as ImageNet [12]) may be difficult and may result in skewed results. Objective evaluation uses tools like computer vision and machine learning to calculate privacy and utility metrics [7]. In this work, we focus on objective evaluation, since these methods are more accurate and scalable than subjective evaluation methods.

2.3. Wireframe Representation of the Human Body

Modern body and shape estimation methods [21, 27, 4, 59, 22, 54] infer a realistic 3D wireframes from a single photo of a person. The estimated wireframe captures realistic body pose and shape of the person. Wireframes, in contrast to silhouette, bounding box, and other avatar representations, provide rich feature maps of the person while removing privacy information. Skinned Linear Person Model (SMPL) [32] is a widely used generative model that represents the human body as a function of shape, pose and translation parameters [29]. In SMPL, pose is represented using the relative rotation of 23 joints, 10 shape parameters are used to represent a person shape and four translation parameters are used to represent global translation. The majority of existing SMPL-based avatar fitting algorithms are solely concerned with inferring SMPL parameters in order to improve shape and pose estimation. To our knowledge, this is the first paper to analyze the privacy and utility aspects of rendered wireframes in comparison to the original video or dataset.

2.4. Video In-Painting

Video in-painting is used to fill the missing regions with plausible content [3]. It is often applied to remove the occluded or unwanted regions in a video sequences [65, 33, 45]. In the early days of in-painting, patch-based methods were used to synthesise missing data by sampling similar spatial and spatio-temporal patches [37, 46]. Deep learning models have made significant progress in video in-painting using recurrent networks [23] and generative adversarial networks [64, 65] to model both temporal and spatial relation of nearby frames in the video. Unlike traditional background modelling methods (correlation-based methods, colour thresholds, and histograms), which extract only static information while ignoring moving objects such as vehicles, video in-painting is used in this work to remove all people from the image while keeping the static and dynamic background.

In this paper, we propose a novel privacy-preserving framework that takes advantage of recent advances in 3D wireframe reconstruction and digital in-painting to create a privacy-enhanced version of a given video. The proposed framework is different from existing literature [7, 40] in two critical ways. Firstly, wireframes are used to represent people in images, allowing for rich features such as people's presence and pose to be represented while maintaining a high level of privacy. Secondly, to the best of our knowledge, we are the first to thoroughly analyze the effects of privacy and utility on visual abstraction and image filtering techniques.

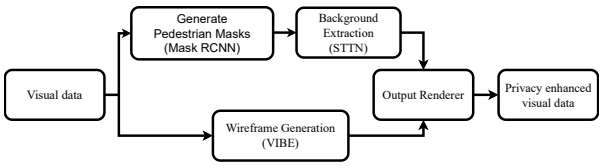


Figure 1: Overview of the Proposed Framework

3. Framework

The goal of the proposed framework is to generate a privacy-enhanced version of an image sequence or video by representing pedestrians with their respective wireframes while preserving their pose, shape, and context in the surrounding scene. To this end, we propose a novel end-to-end framework to generate privacy-enhanced version of a given video or image sequence by leveraging recent advances in visual background extraction and wireframe generation techniques. As shown in Figure 1, the Proposed framework includes three stages: in the first stage, wireframe representations of the pedestrians are rendered; in the second stage, background scene information and objects are extracted; finally, both background information and wireframes are rendered as a final image where pedestrians are represented as wireframes, preserving both the original pose and shape information while offering enhanced privacy.

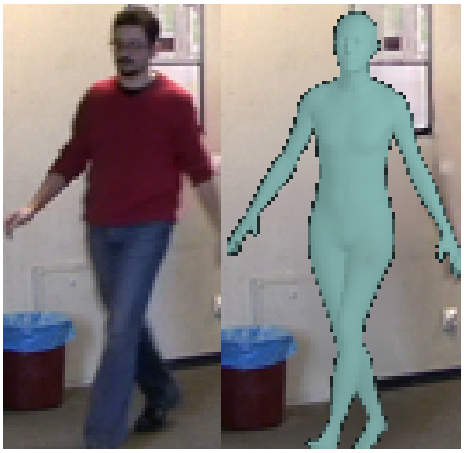
This section describes first how wireframes are generated for a given set of images, and then how the background context information is extracted.

3.1. Wireframe generation

In this paper, the Skinned Linear Person Model (SMPL) [32] is used to represent people as wireframes to protect the privacy of individuals while preserving the statistical information useful for visual analysis. The SMPL model is a generative model that represents the 3D wireframe mesh of the human body as a function of shape, pose. It is defined as $M(\theta, \beta) \in \mathbb{R}^{6890\times 3}$ where $\theta \in \mathbb{R}^{3\times K}$ is the relative rotation of $K = 23$ joints in axis angle representations and $\beta \in \mathbb{R}^{10}$ is the shape's space parameters. The function $M(\theta, \beta)$ generates a triangulated mesh with 6,890 vertices.

The human pose and shape estimation network (VIBE[1]) [27] is used to generate wireframes for pedestrians in given frames. To capture the sequential nature of image sequences, VIBE employs a convolution neural network pre-trained on a single image[21], followed by a temporal encoder and a motion discriminator. For given frames $\{F_t\}_{t=0}^{T}$ with N people, VIBE outputs $\sum_{i=1}^{N}[(P_1^i, P_2^i, ..P_t^i)]$ where $P = [\theta, \beta, \gamma]$ is a vector of the relative rotation, shape, and translation parameters at time step t for the i^{th} person. The translation parameters (γ) are

[1] https://github.com/mkocabas/VIBE



Original | Wireframe representation

Figure 2: Wireframe representation of given image

calculated using a weak perspective camera used to indicate the global translation of 3D mesh.

By representing each pedestrian with a wireframe as depicted in Figure 2, the following privacy information from individuals is removed:

- Hard bio-metric [52] information such as face and other facial features (like hair)
- Soft bio-metric [52] information such as clothing information, which can be used for tracking and re-identification [68]

3.2. Background Extraction

Traditionally, background extraction has been used to separate foreground and background objects. To perform background extraction, the most commonly used approach is to create an explicit model of the background. After that, foreground objects are identified by calculating the pixel difference between the current frame and the background frame [44, 62, 69, 70]. These techniques remove all moving objects from a video or image sequence, leaving only static objects in the scene. In pedestrian analysis, however, moving objects such as vehicles provide critical context information. For this reason, the framework uses a digital in-painting technique to extract the background in order to extract both static and dynamic context information.

Digital in-painting is used to remove occluded or unwanted regions in images or videos for image restoration or enhancement applications. In this work, all pedestrians in image sequences or videos will be removed in order to capture both static and dynamic backgrounds using a pre-trained joint Spatial-Temporal Transformer Network (STTN[2]) [65]. In contrast to other approaches [33, 45], which use only nearby frames to fill missing regions using pixel attention without any temporal coherence, STTN uses multi-scale patch-based attention using both nearby and distance frames. It takes advantage of spatio-temporal redundancies within multiple frames to provide context information while avoiding any occlusions in the foreground. Mask-RCNN[19] is used to generate pedestrian masks that are used to provide input to the STTN network along with the input image frames.

[2] https://github.com/researchmm/STTN

4. Privacy and Utility Metrics

Privacy-preserving techniques provide a varying degree of privacy protection at the expense of loss of data. To quantify this trade-off between privacy and utility, this paper introduces Privacy and Utility metrics, which represent the effectiveness of a privacy-preserving framework in protecting individual privacy while preserving statistical insights for a given pedestrian dataset.

4.1. Privacy Metric

Traditional privacy metrics such as differential privacy [14] are commonly used in private visual datasets, which primarily focus on the leakage of sensitive data from trained models. However, traditional privacy metrics are not suitable to quantify the similarity between the original and privacy-enhanced visual data; instead, they quantify the privacy effectivity of a trained model or inferred data. Subjective evaluation is extensively used in existing literature to evaluate the privacy metric for the original and privacy enhanced data such as user studies to match the privacy protected face images [63], garments, recognisability and shape information[8, 40]. Most of the existing work focus on perceptual understanding of human observers to quantify the privacy but fails to address and evaluate the machine understanding of privacy-enhanced data. In this work, we propose a novel privacy metric by quantifying the similarity between the original and privacy-enhanced data using machine perception.

In pedestrian datasets, privacy protection refers to the ability to avoid disclosure of personal identity information to an adversary. The personal information from images can be used for attribute and identity disclosure [30]. Attributes can be used to infer high-level semantic information about pedestrians [68]. This semantic information can be further utilized in pedestrian tracking, retrieval and re-identification [31, 58, 61]. Identity disclosure refers to linking an individual's information to the record in a database [26]. Any privacy-preserving algorithm should be able to redact both attribute and identity information since one information can be used to uncover another. The wireframe representation of a pedestrian effectively removes all privacy attributes (such as upper body clothing, lower body clothing, hair style etc) as shown in Figure 2. It enables us

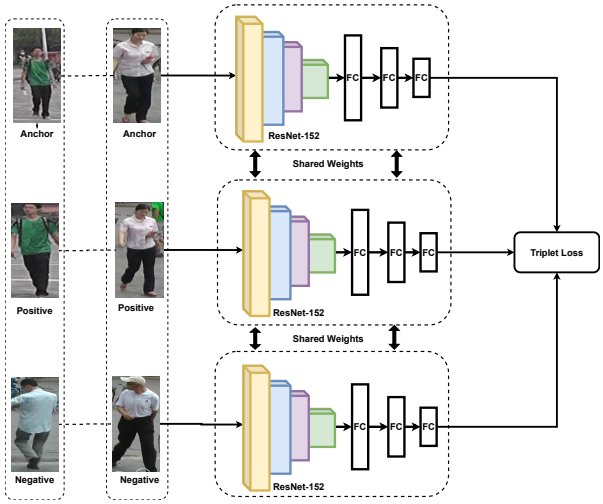


Figure 3: The Architecture of Si-Net

to define privacy solely on the basis of identity discourse. In deep convolutional neural networks, identity is determined by comparing the similarity between the image feature embeddings[57] of original and privacy enhanced images. If the similarity between a pedestrian in a privacy-enhanced image and the original image is low, the level of privacy is considered to be high. The Privacy metric (P_M) solely based on the similarity is given by

$$P_M = 1 - \sum_{i=0}^{N} P_{SI}\left(D_{org}\left[I_i\right], \hat{D}_{org}\left[I_i\right]\right) \quad (1)$$

where P_{SI} is the similarity index of two images and $D_{org}\left[I_i\right]$, $\hat{D}_{org}\left[I_i\right]$ denotes the i^{th} image in the original and enhanced datasets respectively. More information on the similarity index is presented in the following section.

4.1.1 Similarity Index

To understand the effect of a privacy-preserving framework to retain original image information irrespective of privacy attributes, an image similarity index is introduced. The Similarity Index is used to represent the machine perception of image similarity by comparing the machine-learnt representation of the original and privacy-enhanced images.

A Siamese Convolution Neural Network (called Si-Net) is proposed to calculate the similarity index between a pair of images. The architecture of Si-Net is shown in Figure 3. The proposed architecture consist of a Siamese network with three similar convolution neural networks (CNNs) with shared weights. Each CNN consist of pre-trained ResNet-50 along with fully connected layers. The CNNs are used to calculate the feature representation $f(x) \in \mathbb{R}^d$ for a given image i into d dimensional vector space. The network input is triplets of samples consist of

- Anchor Image (I_A)
- Positive Image (I_P)
- Negative Image (I_N)

where the positive image is similar to the anchor image and the negative image is an image unrelated to the anchor. The metric learning is used to train the network using triplet loss. The triplet loss optimizes weights and biases of the model such that the distance between $f(I_A)$ and $f(I_P)$ is bigger than the distance between $f(I_A)$ and $f(I_N)$. The distance between the feature vectors are calculated using the L2-norm as follows

$$D_{A,\vartheta} = \|f(I_A) - f(I_\vartheta)\|^2 \;\; \forall \vartheta \in (P, N) \quad (2)$$

The triplet loss is given by

$$\sum_{i=0}^{N} \left[\|f(I_A^i) - f(I_P^i)\|^2 - \|f(I_A^i) - f(I_N^i)\|^2 + \alpha\right] \quad (3)$$

where N is the total number of triplets and α is the margin enforced between positive and negative pairs.

During inference, the cosine similarity metric is used to calculate the similarity index between the original (I_{org}) and the privacy enhanced image ($\hat{I}_{org}$) as:

$$P_{SI}\left(I_{org}, \hat{I}_{org}\right) = \frac{f(I_{org}) \cdot f(\hat{I}_{org})}{\|f(I_{org})\| \cdot \|f(\hat{I}_{org})\|} \quad (4)$$

4.2. Utility Metric

Pedestrian images contain an abundance of information, such as the identity, appearance, and activities of people. It is essential that the privacy-enhanced dataset can be used for data analysis, and the trade-off for a given privacy-preserving algorithm should be such that the privacy-enhanced dataset can provide required utility while offering the best level of privacy protection. The utility of a dataset is specific to a given application. For example, to analyze pedestrian behavior in crosswalks, video or image sequences in a dataset are used to determine the total number of people using crosswalks and their walking patterns. Walking patterns (such as standing, walking, running, and jogging) are predicted by localizing the keypoints in a person's body. Keypoints are spatial points of the key objet parts of a person in an image. In existing literature, the utility of a dataset was only calculated by counting and detecting pedestrians, and this is only suitable for very limited applications of pedestrian datasets. In this work, a keypoint detector is used along with a pedestrian detector to evaluate the generic utility metric for pedestrian analysis applications. The utility metric is defined as:

$$U_M = \frac{1}{2}\left(F1_{pd} + F1_{kd}\right) \quad (5)$$

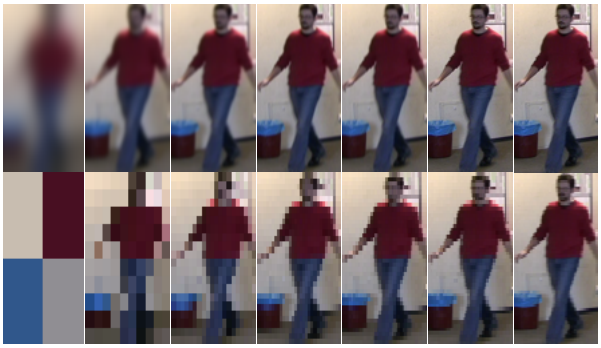


Figure 4: The effect of blurring (top row) and pixelation (bottom row) of the image with varying kernels.

Where $F1_{pd}$ and $F1_{kd}$ are the F1 scores for the pedestrian detector and keypoint detectors. The F1 scores for the pedestrian and keypoint detector are calculated using the following method:

- Run the pedestrian and keypoint detector for the original dataset (D_{org}) and privacy enhanced dataset ($\hat{D}_{org}$)
- Calculate the F1 score by comparing detections of the original and privacy enhanced datasets.

5. Evaluation

This section first introduces the baseline methods, followed by the datasets used, and finally the experimental setup.

5.1. Baseline Methods

In this section, Traditional Image filtering methods are introduced followed by visual abstraction methods.

5.1.1 Image Filtering Methods

Blurring and pixelation are two widely used image filtering methods that distort the images to improve privacy. Blurring is widely used in google street view for example to modify human faces and other sensitive information [15]. The blurring filter applies a Gaussian kernel to the image [50]. Gaussian Kernel utilizes the neighbouring pixels to modify each pixel of the image. Conversely, the pixelation method divides an image into pixel blocks where the kernel's average colour is computed and the resulting colour is assigned to all pixels in that block [41]. Figure 4 shows a blurred (top) and pixelation (bottom) images using different kernels.

5.1.2 Visual Abstraction Methods

In this work, two state-of-the-art visual abstraction methods are used as baseline methods:

- The segmentation[40] method replaces the pedestrian in the images with their representative segmentation masks. Mask-RCNN [51] is used to generate the segementation masks and the generated masks are rendered with background extracted using Video In-painting to generate privacy enhanced image using segmentation.
- The Neural Art [8] method renders the pedestrians images to different style using a neural art algorithm [16]. Image styles are changed using a neural art algorithm and pedestrians are separated from the style image guided by the segmentation masks. These are rendered with the background to generate the final images.

5.2. Experimental Setup

ResNet-50 [20] is used as the backbone for Si-Net. The MARS dataset [67] is used to train the Si-Net by selecting triplets randomly. Each image was resized to 120x40 px (height x width), and represented using 300 features in Si-Net. The margin for the triplet loss(α) used is 0.1 and network is trained for 100 epochs using an adam optimizer with a learning rate of $1e^{-6}$. RCNN [17, 51] and Keypoint RCNN [19] are used as pedestrian detector and keypoint detector.

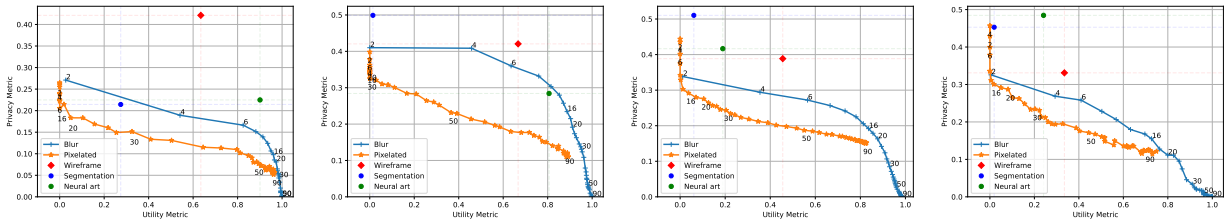
5.3. Datasets

For ease of comparison with the state of the art[40, 7] and to demonstrate the performance on single and multi-person per frame datasets, we run our framework on the PeVID [28], CMU [1], MOT_16 [34] and i.c.sens [38] datasets. The PeVID and CMU datasets contain a single person per frame, while the MOT_16 and i.c.sens datasets consist of multiple people in each frame. The PeVID dataset contains video clips of people performing various actions in both indoor and outdoor settings at different times of the day and night. The CMU dataset is simulated environment data collected for motion capture. In CMU, walking, running and other locomotion videos are considered. The MOT_16 dataset contains real-world video from both static and dynamic camera scenes. The i.c.sens dataset is a collection of pedestrian walking sequences from a road intersection. For a valid comparison of the datasets, only the static camera scenes from the MOT_16 dataset were used in this study.

6. Results

This section presents the results of our evaluation and critically discusses our findings.

Figure 5 presents a comparative analysis of the privacy and utility metrics for the PeVID, CMU, MOT_16 and



(a) PeVID (b) CMU (c) MOT_16 (d) i.c.sens

Figure 5: **Privacy and utility metrics trade-off** for blurring, pixelation, segmentation, neural art, and the proposed framework for different datasets. For blurring and pixelation, the metrics are calculated across a range of kernels (k=2 to 90) to evaluate the relation between privacy and utility. The proposed wireframe method achieves a better trade-off by achieving a higher privacy for the given utility level

Method	PeVID			CMU			MOT_16			i.c.sens		
	$F1_{pd}$	$F1_{kd}$	U_M	$F1_{pd}$	$F1_{kd}$	U_M	$F1_{pd}$	$F1_{kd}$	U_M	$F1_{pd}$	$F1_{kd}$	U_M
Segmentation	0.60	0.04	0.32	0.02	0.01	0.015	0.66	0.01	0.33	0.02	0.01	0.015
Neural art	0.89	0.90	0.89	0.83	0.77	0.8	0.24	0.13	0.18	0.25	0.22	0.23
Wireframe	**0.60**	**0.66**	**0.63**	**0.62**	**0.71**	**0.66**	**0.55**	**0.35**	**0.45**	**0.38**	**0.28**	**0.33**
Blurring	0.61	0.46	0.53	0.73	0.53	0.63	0.67	0.32	0.49	0.54	0.04	0.29
Pixelation	0.24	0.26	0.25	0.56	0.60	0.58	0.56	0.28	0.42	0.32	0.23	0.27

Table 1: Pedestrian detection ($F1_{pd}$), keypoint estimation ($F1_{kd}$) F1 scores and Utility Metric (U_M) values using multiple methods for the PeVID, CMU, MOT_16, and i.c.sens datasets.

i.c.sens datasets. For the traditional blurring and pixelation methods, the privacy and utility metrics are calculated across a range of kernels (from 2 to 90). For both pixelation and blurring, the utility decreases rapidly over the entire range as the privacy increases, and the utility is almost zero when the privacy is greater than 0.3. This confirms that both pixelation and blurring have a severe privacy-to-utility trade-off irrespective of kernels. It can be observed that the proposed framework (labelled wireframe) outperforms both the segmentation and neural art methods by providing an optimal trade-off between privacy and utility without compromising either aspect.

For the single-person-per-frame datasets (PeVID and CMU), the proposed method is able to achieve a higher privacy at a given utility compared to all other models, including traditional blurring and pixelation. The proposed method is able to replace a pedestrian with a wireframe representation, which not only removes all privacy attributes but also present variations in the body shape leading to higher privacy compared to other methods. For multi-person-per-frame datasets, it can be observed that the proposed method achieves overall a better utility than the other methods with respect to privacy. Although the proposed method is able to achieve a good privacy-utility trade-off, in few cases, it still provides lower privacy compared to the segmentation and neural art methods for multi-person-per-frame dataset. For the multi-person datasets, as depicted in Figure 6, partial and full occlusions lead to overlapping of pedestrian representations, which results in a decreased similarity score leading to higher privacy compared to wireframe method. We can also notice that due to occlusions, the utility metrics are far lower compared to the wireframe methods. In addition, we can also note that wireframe methods ignore pedestrians if there are occlusions, which leads to a low utility metric.This limitation will be addressed in future work.

Figure 6 shows few of the privacy enhanced images using different methods for different datasets. As depicted in Figure 6, segmentation method offers a shallow representation of the pedestrians, whereas neural art method representation hides the pedestrian by style transfer. In contrast to segmentation and neural art, the proposed method offers a better individual representation using wireframes. Table 1 presents the pedestrian and keypoint detector F1 scores. To compare the effectiveness of the proposed method, the kernels for both blurring and pixelation are selected to match the wireframe utility metric value. The kernel values selected for the PeVID, CMU, MOT_16 and i.c.sense datasets are (36,60,46,46) and (4,5,4,4) for pixelation and blurring respectively. When compared to all other meth-

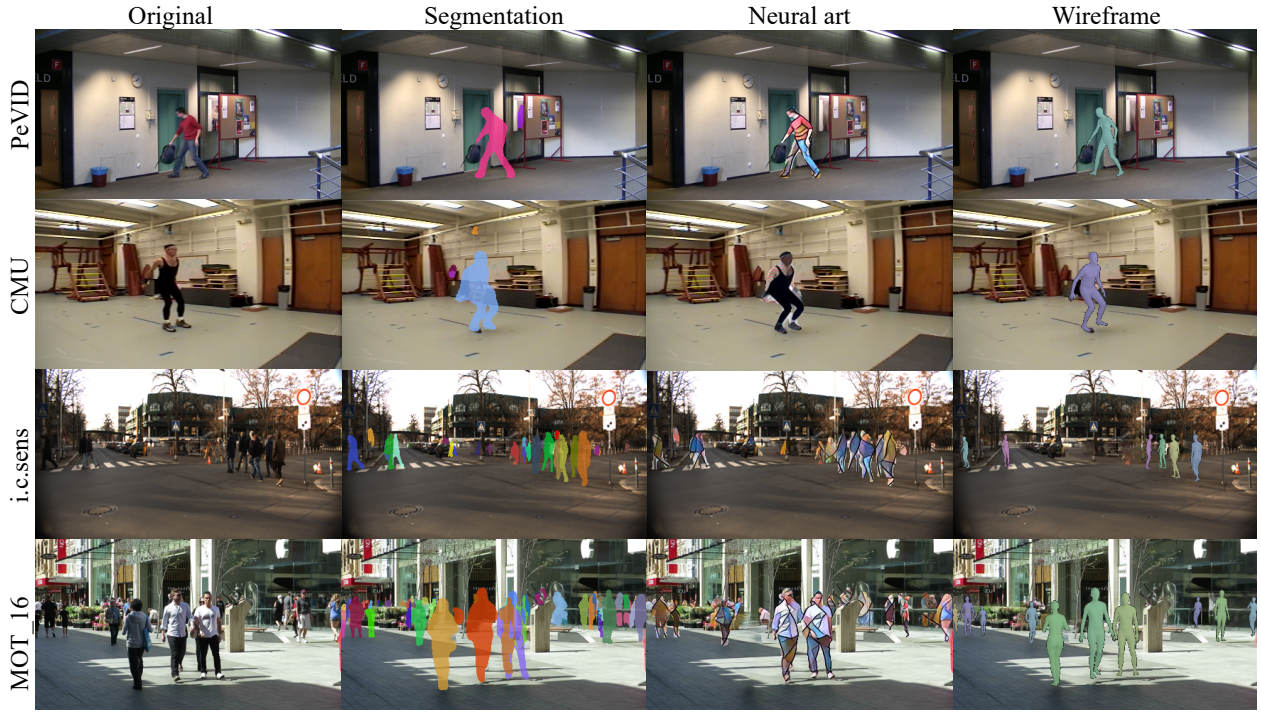


Figure 6: Qualitative results for the proposed (Wireframe) and baseline methods (Segmentation and Neural art)

ods, the proposed method is able to achieve a better performance for keypoint detectors. This is due to the wireframe model's ability to represent poses as well as a person's shape. The performance of keypoint and pedestrian detectors in multi-person-per-frame datasets is significantly decreased for both segmentation and neural art methods. This may be due to void representation of segmentation and occlusions in neural art base representations as shown on the qualitative results in Figure 6. However, the proposed wireframe method is capable of achieving a performance comparable to traditional methods while also providing enhanced privacy metric.

7. Conclusion and Future Work

This paper presents a privacy-preserving framework that generates a privacy-enhanced version of a given video or image sequence to enable privacy-preserving pedestrian behavior analysis. In contrast to existing visual abstraction privacy-preserving methodologies, pedestrians are represented using a quantitative wireframe to improve privacy while maintaining the pose and shape of the pedestrian. Two generic metrics are introduced to evaluate respectively the privacy and the utility of a given video or image sequence. The proposed framework is able to outperform existing state of the art visual abstraction methods by providing an improved privacy for the same utility. The results show that the proposed framework is able to achieve a better privacy utility trade-off compared to the existing state-of-the-art methods by improving the utility of privacy-enhanced datasets. However, the proposed framework still has its limitations in regards to how much privacy can be enhanced and how well it compares with other methods. Additionally, the proposed privacy metric does not consider gait parameters, which may be exploited to compromise the individual's privacy.

Future work includes enhancing the privacy metric score for the wireframe representation by exploiting the SMPL shape parameters. The wireframe method supports dynamic changes in the body and shape of pedestrian which enable us to investigate gait patterns, while preserving privacy along with body shape variations. It would also be interesting to experimentally compare the subjective and similarity metric-based privacy evaluation methods.

Acknowledgement

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

References

[1] Cmu graphics lab motion capture database.

[2] Ben Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. *CVPR 2011*, pages 3457–3464, 2011.

[3] M. Bertalmio, A.L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016.

[5] Péter Borosán, Ming Jin, Doug DeCarlo, Yotam Gingold, and Andrew Nealen. Rigmesh: automatic rigging for part-based shape modeling and deformation. *ACM Transactions on Graphics (TOG)*, 31(6):1–9, 2012.

[6] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10, 2000.

[7] Karla Brkić, Tomislav Hrkać, and Zoran Kalafatić. Protecting the privacy of humans in video sequences using a computer vision-based de-identification pipeline. *Expert Systems with Applications*, 87:41–55, 2017.

[8] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328, 2017.

[9] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. The wildtrack multi-camera person dataset. *ArXiv*, abs/1707.09299, 2017.

[10] Yin Chen, Gang Dang, Zhi-Quan Cheng, and Kai Xu. Fast capture of personalized avatar using two kinects. *Journal of Manufacturing systems*, 33(1):233–240, 2014.

[11] Kenta Chinomi, Naoko Nitta, Yoshimichi Ito, and Noboru Babaguchi. Prisurv: Privacy protected video surveillance system using adaptive visual abstraction. In *International Conference on Multimedia Modeling*, pages 144–154. Springer, 2008.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[13] Frederic Dufaux and Touradj Ebrahimi. Scrambling for privacy protection in video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(8):1168–1174, 2008.

[14] Cynthia Dwork. Differential privacy. ICALP'06, page 1–12, Berlin, Heidelberg, 2006. Springer-Verlag.

[15] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*, pages 2373–2380. IEEE, 2009.

[16] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *Journal of Vision*, 16(12):326–326, 2016.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[18] Jules. Harvey, Adam. LaPlace. Exposing.ai, 2021.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[22] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[23] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5792–5801, 2019.

[24] Jinsu Kim and Namje Park. A face image virtualization mechanism for privacy intrusion prevention in healthcare video surveillance systems. *Symmetry*, 12(6):891, 2020.

[25] Itaru Kitahara, Kiyoshi Kogure, and Norihiro Hagita. Stealth vision for protecting privacy. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 404–407. IEEE, 2004.

[26] Koung-Suk Ko, Woo-Jin Ahn, Geon-Hee Kim, Myo-Taeg Lim, Tae-Koo Kang, and Dong-Sung Pae. Re-identification for multi-object tracking using triplet loss. In *2021 International Conference on Information Networking (ICOIN)*, pages 525–527, 2021.

[27] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020.

[28] Pavel Korshunov and Touradj Ebrahimi. Pevid: privacy evaluation video dataset. In *Applications of Digital Image Processing XXXVI*, volume 8856, page 88561S. International Society for Optics and Photonics, 2013.

[29] Anil Kunchala, Mélanie Bouroche, Lorraine D'Arcy, and Bianca Schoen-Phelan. Smpl-based 3d pedestrian pose prediction. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.

[30] Wenjing Liao, Jianping He, Shanying Zhu, Cailian Chen, and Xinping Guan. On the tradeoff between data-privacy and utility for data publishing. In *2018 IEEE 24th International*

Conference on Parallel and Distributed Systems (ICPADS), pages 779–786, 2018.

[31] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[33] Yasuyuki Matsushita, Eyal Ofek, Weina Ge, Xiaoou Tang, and Heung-Yeung Shum. Full-frame video stabilization with motion inpainting. *IEEE Transactions on pattern analysis and Machine Intelligence*, 28(7):1150–1163, 2006.

[34] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[35] Carman Neustaedter and Saul Greenberg. The design of a context-aware home media space for balancing privacy and awareness. In *International Conference on Ubiquitous Computing*, pages 297–314. Springer, 2003.

[36] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(1):1–36, 2006.

[37] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez. Video inpainting of complex scenes. *Siam journal on imaging sciences*, 7(4):1993–2019, 2014.

[38] Uyen Nguyen. I.c.sens multi-view pedestrian tracking dataset, Jul 2020.

[39] V. A. Niţă and V. Popa. A framework for privacy assurance in a public video-surveillance system. In *2019 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4, 2019.

[40] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2018.

[41] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. Visual privacy protection methods: A survey. *Expert Systems with Applications*, 42(9):4177–4195, 2015.

[42] Sangho Park and Henry A Kautz. Privacy-preserving recognition of activities in daily living from multi-view silhouettes and rfid-based training. In *AAAI Fall symposium: AI in eldercare: new solutions to old problems*, pages 70–77, 2008.

[43] Sangho Park and Mohan M Trivedi. A track-based human movement analysis and privacy protection system adaptive to environmental contexts. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, pages 171–176. IEEE, 2005.

[44] Donovan H Parks and Sidney S Fels. Evaluation of background subtraction algorithms with post-processing. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 192–199. IEEE, 2008.

[45] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmio. Video inpainting of occluding and occluded objects. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–69. IEEE, 2005.

[46] Kedar A Patwardhan, Guillermo Sapiro, and Marcelo Bertalmío. Video inpainting under constrained camera motion. *IEEE Transactions on Image Processing*, 16(2):545–553, 2007.

[47] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[49] Kenny Peng. Facial recognition datasets are being widely used despite being taken down due to ethical concerns. here's how., Oct 2020.

[50] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living. *arXiv preprint arXiv:2112.09422*, 2021.

[51] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[52] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016.

[53] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshops*, 2016.

[54] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[55] Archana Sapkota and T. Boult. Large scale unconstrained open set face database. *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013.

[56] Jake Satisky. A duke study recorded thousands of students' faces. now they're being used all over the world, Jun 2019.

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[58] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*, pages 475–491. Springer, 2016.

[59] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human mesh recovery from monocular images via a

skeleton-disentangled representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5349–5358, 2019.

[60] Carlo Tomasi. Letter: Video analysis research at duke, Jun 2019.

[61] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

[62] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):780–785, 1997.

[63] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pages 25313–25330. PMLR, 2022.

[64] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

[65] Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *European Conference on Computer Vision*, pages 528–543. Springer, 2020.

[66] Cha Zhang, Yong Rui, and Li-wei He. Light weight background blurring for video conferencing applications. In *2006 International Conference on Image Processing*, pages 481–484. IEEE, 2006.

[67] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[68] Jianqing Zhu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Z. Li. Pedestrian attribute classification in surveillance: Database and evaluation. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 331–338, 2013.

[69] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 28–31. IEEE, 2004.

[70] Zoran Zivkovic and Ferdinand Van Der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006.