ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.33, No-1-2, 2025

# Preliminary Study of a Non-Direct Generative Image Anonymization Pipeline for Anomaly Detection

Ivan Nikolov
Department of Architecture, Design and Media Technology
Aalborg University, Rendsburggade 14, Denmark, Aalborg
iani@create.aau.dk

## ABSTRACT

With growing General Data Protection Regulation (GDPR) compliance demands for deep learning surveillance models, human anonymization is a key research area. Most studies use RGB images as input for generative models, which retain demographic features, compromising anonymization and consistency across frames. We present our initial study into a full-body anonymization pipeline for anomaly detection datasets, where the synthetic person generation never has access to the RGB pedestrian visuals. The proposed pipeline uses a combination of existing models for easier reproducibility. We use YoloV8 for object detection, ClipSeg and BiRefNet for segmentation, OpenPose for pose estimation, and an animation diffusion model. The diffusion model processes only masks and skeletal pose images, removing the problems with using sensitive data. We test on the Avenue dataset. We show that the proposed pipeline can consistently anonymize and change the demographics of detected pedestrians. We discuss the observed problems and the next steps in building a more robust second version.

## Keywords

anomaly detection, synthetic data, generative models, open-pose, anonymization, surveillance

## 1 INTRODUCTION

In deep learning, capturing, analyzing, and storing data containing individuals' likenesses without proper consent or anonymization can lead to non-compliance with GDPR laws [VVdB17]. To address this, traditional anonymization techniques such as pixelization, blurring, or blacking out facial features, and even full-body pixelization, have been used. However, applying these methods too aggressively risks distorting and corrupting the data, leading to poorer performance in training deep learning models. Conversely, insufficient blurring or pixelization may be reversible [XXH*21], potentially revealing an individual's identity.

A more powerful way for dealing with privacy preservation is by employing deep neural anonymization, for generating facial and body overlays with better results [HL23b]. Most of these methods focus on facial anonymization [KLY*21, HML19], which leaves the problem of identity detection through secondary features like clothing, body type, gender, etc. With the stricter GDPR requirements for anonymization, more models focus on full-body anonymization [HSML23, HL23a, MELT20]. These methods employ surface-guided GANs, which utilize dense pixel-to-surface correspondences, as well as shape and face priors for controlling the identity of the generated synthetic augmentations. Diffusion-based generative models have been used for face anonymization with better results than GANs [HZC*23, PBLM24]. A problem with both GAN and diffusion models is that they directly use RGB data, which can expose them to reverse engineering from malicious actors for data gathering [LXW*20].

Privacy preservation is especially important in the field of surveillance anomaly detection. Outdoor scene anomaly detection is a widely researched field utilizing semi-supervised learning methodologies for detecting anomalous frames signaling of anomalous behavior, emergencies, and unexpected scenarios, from a stream of normal expected ones. For this, large amounts of video and image data are captured from public areas, featuring many humans, who have not provided their consent to the data capture. Concerns for privacy preservation have resulted in research into data unlearning [FWZ*22] and ways to make anomaly detection less invasive through semantic segmentation [BDNM21].

In this work, we explore the possibilities of full-body anonymization for anomaly detection datasets using a diffusion-based pipeline and the effects of anonymization on anomaly detector models (Figure

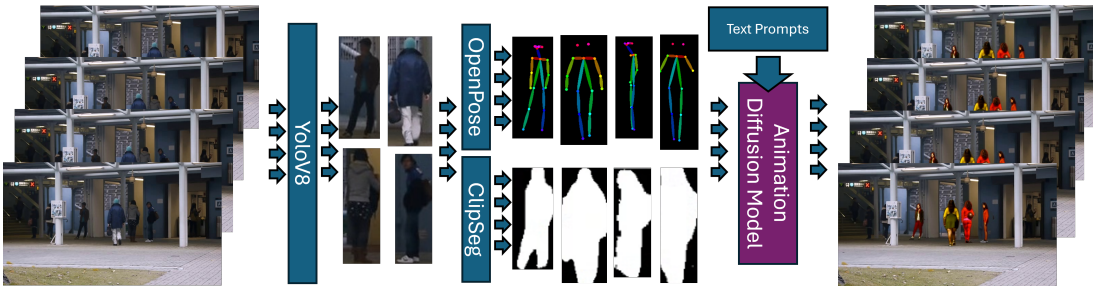


Figure 1: Overview of the proposed full-body anonymization pipeline. Pedestrians are detected and tracked in the input frames using YoloV8 with BoT-SORT, and the bounding box images are segmented using ClipSeg to extract the pedestrian masks and reconstruct the backgrounds. OpenPose is used to extract the skeletonization of the detected pedestrians, which is given together with the masks to a diffusion-distillation model, together with randomized appearance prompts to generate the anonymized images.

Figure 2: Examples of anonymized frames using the proposed pipeline, where the generative model only uses skeletonizations and masks.

1). The proposed solution takes inspiration from other selective generative humans for anonymization research [SSS*24] but relies on existing solutions for making it easier to reproduce. employs the YoloV8 model [JCQ23], combined with the ClipSeg model [LE22] and OpenPose [CSWS17] to detect pedestrians and capture their poses. We concatenate multiple tracked frames from each pedestrian and feed them to a diffusion-distillation model [LY24] which can produce animation sequences (Figure 2). We test our proposed pipeline on the Avenue dataset [LSJ13]. We compare the anonymized augmented images to traditional anonymization techniques, as well as state-of-the-art deep neural anonymization models. We train four anomaly detection models on the produced datasets. We show that the proposed anonymization pipeline, which does not give the generative model access to the RGB human features, produces frame-consistent results. The main contributions of this paper can be summarized as:

- Building a pipeline for creating context-specific and multi-frame consistent full-body anonymization without giving RGB information directly to the generative model;
- Demonstrating the effects of different types of anonymization techniques on anomaly detection models;
- Showing that our proposed pipeline produces augmentations that produce a small amount of interference with the performance of anomaly detectors while being able to completely anonymize the demographic.

The code for the proposed pipeline can be found on GitHub - `https://github.com/IvanNik17/Generative-Image-Anonymization-Pipeline`.

## 2 RELATED WORK

### 2.1 Image People Anonymization

Initial work has focused on simple identity anonymization employing naive approaches like facial black bars, blurring, pixelization, and distortion. These methods can work well for simpler cases where the subjects

do not move a lot and only parts of them are visible - for example for facial-only anonymization from the torso up [RD16]. The problem with these approaches is that they do not provide enough anonymization for secondary characteristics. Features like clothing, body shape, and gender, among others, can be used for identifying people and to anonymize them. Thus, stronger pixelization or blurring needs to be applied. But these could also result in the introduction of noise and artifacts that make using the data more difficult or, in some cases, impossible. There has been new research in making these blurring and pixelization approaches more dynamic and better optimized towards obscuring only certain features [AJO24].

Later work focused more on facial anonymization using facial-identity preserving features and k-facial attributes [RLR18]. These works provide better overall anonymization, but add additional noise to the images, which can be detrimental when used for training other models. More recent works focus on deep neural anonymization using generative models, which transform the detected underlying people [RLR18] or segment and inpaint them [KLY*21, HML19]. From these methods, a limited number also focus on full-body anonymization with most producing artifacts [MELT20, HSML23], while others like the DeepPrivacy2 [HL23a] provide much cleaner results for both full-body and facial anonymization. This method utilizes a modified U-Net architecture with an encoder and decoder modeled around a surface-guided GAN architecture. The encoder and decoder parts have increased depth, and for the full-body anonymization, the model is trained on a specifically curated Flickr Diverse Humans (FDH) dataset consisting of 1.53M annotated human images. This provides it with the capability to generate anonymizations for diverse human poses and overlapping. This also provides a possible problem with the proposed solution. The architecture is trained using the FDH subset from the larger YFCC100M dataset and directly utilizes human images. This, together with the model, needs to be given RGB images to generate the anonymizations that could result in privacy preservation non-compliance. Anomaly detection datasets present additional challenges. For example, datasets like Surveillance Videos [SCS18] are often curated using data from platforms like YouTube and LiveLeak. Combined with research findings that models trained on images containing humans can be exploited to retrieve sensitive information, such as faces [LXW*20], this highlights the need for approaches that minimize privacy risks in anomaly detection models. Furthermore, many anomaly detection datasets are captured at specific times and locations and focus on a limited set of individuals, which can lead to racial or gender bias. It has been shown that unsupervised models trained on datasets like ImageNet can learn such biases [SC21], and there are significant differences in reconstruction errors based on the demographic groups present in the images [BGU22].

### 2.2 People Anonymization Through Diffusion Models

Diffusion models have become much more adept in generating realistic humans, through the use of physics-based modeling [YSI*23] or skeleton-based priors [JZZ*23] for creating more stable synthetic data between multiple frames. This, combined with the facial generation [SVH*24] and animation [ZLG*23] results, containing changes like blinking, facial expressions, and different head tilts, provides a strong foundation for building a realistic animation system. Some have also been adapted to work with anonymization tasks, where the input images are used as prior information to guide the generation task [HZC*23], together with additional prompt information like names [SZS24]. These methods are mostly focused on facial anonymization and use the input image as a guide, thus directly using private information, which can be problematic. Diffusion-based models like CAMOUFLaGE-Base [PBLM24] also outperform GAN-based models like DeepPrivacy2 on re-identification and downstream tasks. The choice of diffusion models for this paper was driven by their higher quality, the limited research on using them for full-body anonymization, and the need for a better method to generate synthetics without directly observing the potentially privacy-breaking input data.

## 3 METHODOLOGY

### 3.1 Pedestrian Tracking and Segmentation

As an initial step, we need to detect the pedestrians, track them between frames, and correctly segment them. For object detection and tracking, we have selected YoloV8 [JCQ23] together with the BoT-SORT tracking algorithm, as it has been shown to perform well with pedestrian tracking [dONRM24]. We use the base version of YoloV8, together with changes to the tracking settings - the lower threshold for second associations between tracklets is set to 0.20, the threshold used for starting new tracks is set to 0.70, and the threshold for matching tracklets is set to 0.70. These values were selected from the best practices outlined in the YoloV8 documentation for more stable tracking, especially in the case of partial occlusions.

Next, we use ClipSeg [LE22] on the extracted bounding box sequences. We choose ClipSeg, as it works well with segmenting lower detail objects and it provides comparable results to SAM [KMR*23] or Yolact

Figure 3: Examples of detected pedestrians, together with the ClipSeg segmentations and the reconstructed backgrounds.)

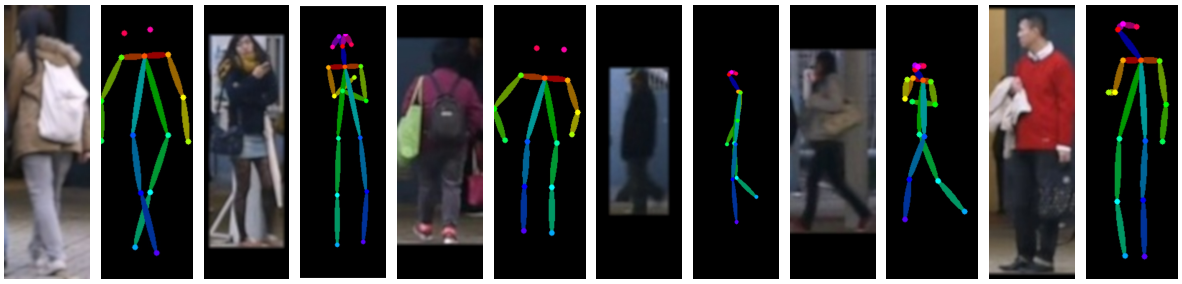

Figure 4: Examples of frames of tracked pedestrians from YoloV8 and extracted pose skeletonizations from OpenPose.

[ZOL*22]. We use the model with a dimensional reduction of 64. We input the extracted bounding boxes, together with a prompt as input to the model, and generate a mask using a sigmoid function. We save the masks and use them to segment the pedestrian from the background. We then reconstruct the full background by blending the image with a background created by applying a temporal median filter on the frames from the sequence. This, of course, means that the current solution is useful for static camera positions. Possible solutions to this are discussed in the Results section. Example results from the segmentation, together with the input pedestrians and reconstructed background, can be seen in Figure 3.

## 3.2 Skeleton Extraction

Once we have extracted the tracked pedestrian bounding boxes, we resize them using cubic interpolation to scale them to a size above the minimum proposed by OpenPose documentation [CSWS17] of 656x368. This step is necessary as some detected pedestrians can be very small, resulting in incorrect skeletonization. We use the pre-trained production-ready OpenPose weights provided as part of ControlNet. In some scenes, detected pedestrians might overlap, and to prevent the skeletonization of multiple people in a single bounding box, the extracted skeletonized bodies are filtered, and only the ones with the highest scores are left, discarding anything else found in the bounding box. The generated sequence skeletonization images are saved for each tracked pedestrian. Results from the OpenPose together with the input bounding box images can be seen in Figure 4. By feeding these skeletonizations to the diffusion model later, we obscure any distinctive features about captured pedestrians' facial appearance, clothes, or demography.

## 3.3 Diffusion Pedestrian Generation

The created skeletonization image sequences from OpenPose are then further filtered to remove broken sequences. This is done by calculating the number of detected skeletons for all images in the sequences and their average OpenPose scores. Each sequence below the specified thresholds of $N * 0.2$ for missing images and 0.35 for the OpenPose score, where $N$ is the number of frames in a sequence, is discarded. This is done as too many fully missing skeletonization images in sequences or sequences with badly estimated skeletons - missing hands, feet, heads, etc., result in unstable diffusion generation of animations. In the current implementation, these pedestrians are not anonymized, but they can also be fully removed by directly reconstructing the background.

The skeleton sequences and segmentation masks are then given to the diffusion-distillation model [LY24], with the variational autoencoder from Stable Diffusion [RBL*21], built through the Diffusers library [vPPL*22]. The model is selected, as it combines and distills the probability flow of multiple base diffusion models into a unified motion module, which helps

-Dark-skinned woman with long hair, in a skirt and high-heel shoes.
- Dark-skinned man with short hair, in a suit and snickers.
- Caucasian non-binary person with medium hair, in a jacket, jeans, and snickers.

- Dark-skinned woman with long hair, in a sundress and shoes.
- Caucasian man with short hair, in a suit and boots.
- Asian woman with short hair, in a hoodie, jeans, and sneakers.

Figure 5: Examples of generated results given three different prompts (shortened to fit in the space) starting with a detected pedestrian, followed by the extracted pose skeletonization and the three results before they are masked and blended in the frame

generate a wider array of styles and variations. This architecture also provides a much better inference efficiency, especially when generating animations with multiple frames. This is especially important for our use case as we want to create longer sequences of multiple pedestrians present in each dataset sequence. Along with the image inputs, we generate a randomized sequence of positive prompts. It has been shown that generative models can exhibit inherent biases related to gender, race, and appearance, which may be amplified by simple or vague prompts [BKD*23]. To mitigate this, we ensure our randomized prompts are highly descriptive, following a specific order: starting with *ethnicity*, then *gender*, followed by secondary characteristics like *body type* and *height*, then *hair color and type*, followed by *occupation*, and finally *clothing*. To emphasize different aspects of the prompts and ensure monochrome backgrounds, we use the Compel library [Ste] for efficient prompt embedding generation. Examples of input images, skeletonizations, and generative model outputs with varying prompts are shown in Figure 5. As the generated synthetic humans have background, we utilize the BiRefNet [ZGF*24] framework for high-resolution segmentation. We utilize this model for segmentation, as it produces exact segmentation, even for smaller details like hair, clothing, and accessories. The model works well with the higher resolution generated pedestrians to produce tight-fitting masks. The masked-out generated humans are then blended into the reconstructed backgrounds.

# 4 EXPERIMENTS

The experimental setup consists of the selected anomaly detection CUHK Avenue dataset as a basis. We select a subset of the dataset and process it with three classical anonymization methods, together with the state-of-the-art GAN-based DeepPrivacy2 model and our proposed diffusion-based pipeline. For testing the quality of the generated anonymizations, we use the produced data to train four anomaly detection models and observe the changes in their performance, together with more qualitative visualization of the output anonymized images.

## 4.1 Anonymization Comparisons

To do an initial testing on our proposed pipeline, we have selected the CUHK Avenue dataset [LSJ13], as it is comprised of normal and abnormal images focused only on pedestrians without vehicles. It also contains pedestrians at different scales and larger groups of people. For these initial experiments, due to hardware limitations, we have decided to generate the anonymous training data from the first 90 frames of each sequence and to compute the animated anonymized sequences from each 30 frames. Together with our pipeline, we use three traditional ways for anonymization - facial masking, full-body Gaussian blurring, and pixelization. Finally, we use the state-of-the-art DeepPrivacy2 model [HL23a], to generate full-body deep neural anonymization. Examples of input pedestrian sub-images together with the generated anonymizations are given in Figure 6.

## 4.2 Anomaly Detection Models

To test how much these anonymizations affect the results of anomaly detection models we train four detectors using the provided datasets - $\text{MNAD}_{\text{pred}}$ / $\text{MNAD}_{\text{recon}}$ [PNH20], LGN-Net [ZLL*22], MPN

Table 1: The frame level AUC results from the five tested anomaly detection models, when trained on the full non-anonymized Avenue dataset ($NA_{full}$), the non-anonymized subset ($NA_{subset}$), as well as the subset anonymized through the three classical techniques - face obscuring, full-body blurring and pixelization, DeepPrivacy2 model and our proposed pipeline. Models are separated into using directly the RGB images, versus using only the skeletonization for generating the anonymization, as well as into simple and neural anonymization. Values are between 0 and 100 and higher values are better.

| Model | No Anonym. | | Simple Anonym. | | | Neural Anonym. | |
|---|---|---|---|---|---|---|---|
| | $NA_{full}$ | $NA_{subset}$ | Obscured | Blurred | Pixelated | DeepPrivacy2 | Ours |
| | | | ♦ Direct RGB | | | ♦ Direct RGB | ♣ Indirect Skeleton |
| $MNAD_{pred}$ | 88.5 | 85.3 | 84.4 | 80.1 | 81.4 | 85.6 | 85.7 |
| $MNAD_{recon}$ | 82.8 | 80.7 | 80.6 | 76.7 | 79.4 | 80.4 | 80.2 |
| MPN | 89.5 | 81.6 | 81.7 | 78.2 | 80.7 | 81.0 | 77.1 |
| LGN-Net | 89.3 | 81.2 | 82.7 | 78.8 | 79.2 | 79.9 | 75.6 |
| LNTRA | 84.6 | 80.2 | 80.4 | 77.8 | 78.6 | 80.2 | 80.3 |

[LCC*21], and LNTRA [AZLL21]. For the MNAD model predictive and reconstructive versions - $MNAD_{pred}$ and $MNAD_{recon}$, we set the separation and compactness to 0.1 and 0.01. Both are trained with a batch size of 4 and a learning rate of $2e-4$. For MPN, we use a batch size of 4 with a learning rate of $1e-4$. We set the frame and feature reconstruction loss weights to 1 and the feature distinction loss weight to 0.1. For LGN-NET, we change the compactness weight to 10 and the separation loss weight to 5, and we set the batch size to 4, and the learning rate to $2e-4$. For LNTRA, we use skip frames and pseudo anomaly in-painting of 0.2, jumps at $3,4,5$, and trained with a learning rate of $1e-4$. All models are trained for the recommended 60 epochs, except MPN, which is trained for 100 epochs. Each of the models is trained for the recommended number of epochs and using the hyperparameters proposed by their respective papers. For evaluation, we compute the frame-level AUC.

## 5 RESULTS

The results from the experiment are given in Table 1. We computed the performance of the four anomaly detector models trained both on the subset of non-anonymized data and took the performance for the full training dataset ($NA_{full}$) as a base comparison from the proposed papers. We can see that even though all four models get a performance hit from using the smaller subset of training data, the difference is not major and is consistent. Thus, we can use the non-augmented subset ($NA_{subset}$) as a baseline. We can see that by obscuring the faces of pedestrians we get a minimum change in performance. This is expected as faces are just a small part of the captured pedestrians and thus hiding them does not hinder the performance of anomaly detectors. The problem with just obscuring the faces is that it does the bare minimum for privacy preservation and leaves many easily identifiable features like clothes, body shape, hands, skin tone, etc. unobscured. The two more destructive anonymizations using blurring and pixelization of the detected pedestrians result in a more noticeable performance drop, as too many of the pedestrians' features are obscured.

DeepPrivacy2 and our pipeline both fully anonymize pedestrians, while taking less of a performance hit than the traditional methods. However, our solution often lags behind DeepPrivacy2 and can show artifacts from the blending process. This is expected, as our approach separates human detection from synthetic anonymization using a diffusion model. The ControlNet-based generative model relies only on OpenPose skeletons and ClipSeg masks, without ever seeing the detected pedestrians or privacy-compromising RGB data. This limits its blending and color correction capabilities compared to DeepPrivacy2. However, our method could offer better privacy compliance by allowing pedestrian tracking and segmentation on sensitive RGB data to be handled separately. MPN and LGN-Net models are most affected by performance drops due to their sensitivity to frame changes, which occur when OpenPose and Yolo detections fail, leading to inaccurate generations. It's interesting to point out that $MNAD_{pred}$ even gives better results when using our anonymization, likely because it provides more variable data, which helps with the overfitting on the smaller subset, as the model differs from the others with predicting what the next frames will be. We can also show that even though the imperfect blending is creating background artifacts in the proposed pipeline, it generates temporally consistent frames similar to DeepPrivacy2, with fewer details popping in between consecutive frames (Figure 7). To test this, we gather detected sequences of pedestrians through Yolov8 and extract bounding boxes from the non-anonymized images, the DeepPrivacy2 anonymized images, and the ones anonymized by our pipeline. We then calculate the SSIM metric between each two images in a sequence. Higher values would signify that sequential pedestrian

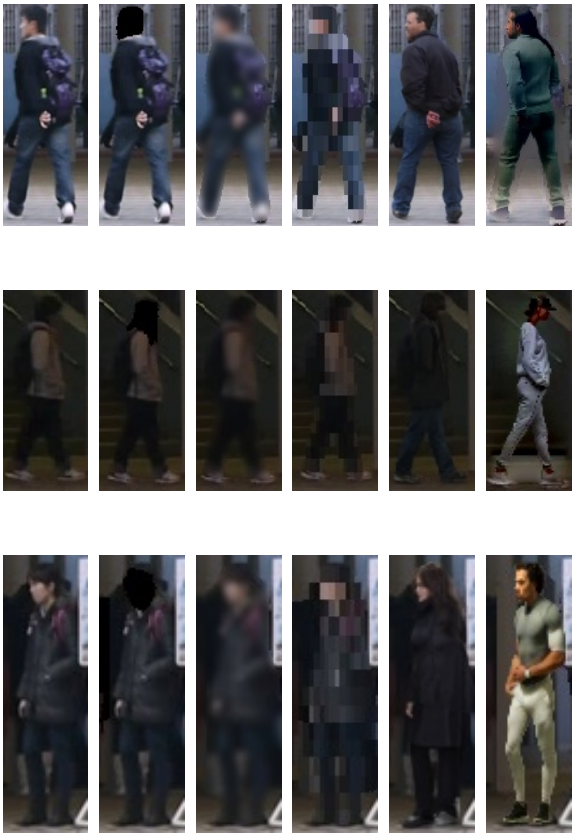

Figure 6: Examples of the different anonymization techniques used for the anomaly detection training data. From left to right - base image, obscured, blurred, pixelized, anonymized using DeepPrivacy2, and finally ours. Because of the blending when we have a darker background, we can see that our solution lightens it.

anonymizations are more consistent, as humans do not change drastically between frames. The results in Table 2 show that both models have consistency degradation, with our pipeline's consistency suffering when the pedestrian detection or pose skeletonization breaks. We can see that we achieve similar performance to DeepPrivacy2, in full-body anonymization, without the need to provide a full RGB image to the generative model, providing better privacy preservation at the cost of visual artifacting. Our model has inconsistency with blending artifacts, incorrect detections, and separating overlapping people, but provides better temporal consistency.

# 6 PROBLEM DISCUSSION

Our initial research into using diffusion models for full-body anonymization shows positive results that can make more diverse fully anonymized data that is reactive to changes in the images. However, as this is a work in progress, some problems need to be addressed

Figure 7: Examples of the temporal consistency of DeepPrivacy2 (top row) compared to our pipeline (bottom row). Our results have some background artifacts from the blending, but generate more temporally consistent anonymizations.

Table 2: Results from running a pairwise SSIM calculation on pedestrian sequences to determine the temporal consistency of the calculated anonymizations.

| **Dataset** | **Avg. SSIM Pairs** |
|---|---|
| Non-Anonymized | 0.895 |
| DeepPrivacy2 | 0.810 |
| Ours | 0.818 |

before the pipeline can be useful in larger projects. In this section, we will discuss these and give our recommendations, together with showing examples of failure cases, based on tracking and pose estimation.

## 6.1 Performance and Speed

Even with using a highly optimized diffusion model, the required resources for generating larger frame sequences become very high quickly. We provide detected OpenPose skeletonization images resized above the base suggested resolution for good results of 512x512. We synthesized the anonymized pedestrians on a PC with an Nvidia RTX 4070 Ti Super, 32 GB of RAM, and an Intel i7-13700KF. In this configuration, it was shown that the provided 12GB of VRAM was not enough for generating sequences larger than 30 frames before crashes or visual artifacts started to appear. In addition, the generation times for the full pipeline were measured between 10 minutes and 42 minutes, depending on how many pedestrians were present through the provided image sequences. In datasets with larger resolutions, these resource requirements and processing times would also grow. It would be useful to experiment with the sizes of the generated

(a) (b) (c)

(d) (e) (f)

Figure 8: Examples of anonymized frames with incorrectly generated humans, caused by errors in pedestrian detection, pose skeletonization, or too large masks.

sequences to strike a balance between generation times and performance.

### 6.2 The Reliance on Correct Pedestrian Detections

The generative parts of our pipeline rely on correct inputs from OpenPose, which in turn rely on good inputs from object detection, tracking, and segmentation. A cascade of failures can arise, where incorrectly detected bounding boxes from YoloV8 result in incomplete segmentation from ClipSeg, and OpenPose results in missing parts of the body. Which in turn is input to the diffusion model that fails to produce any usable results, by hallucinating incorrect shapes. Especially when people get obscured by the foreground or by other people, and the tracker incorrectly switches. A way to address this is by exploring better solutions for tracking in crowded environments [SDABMP21]. Another possibility is refining the OpenPose results to better take into account consecutive detected frames in denser environments with overlap [WZH21] through the use of Pose-based Inference and part-based Pose generation. Examples of incorrectly generated anonymizations can be seen in Figure 8. These are caused by incorrect pedestrian detection and tracking by YoloV8 or by incorrectly calculated skeletonizations from OpenPose. This incorrect data is then propagated to the generative model and results in unnatural poses (Figure 8a), blurry (Figure 8b), or failed generations (Figure 8d).

### 6.3 Better Background Blending

Our pipeline currently utilizes a simple blending step between the generated pedestrian and the reconstructed background, which utilizes the masks calculated by BiRefNet. The background's reconstruction uses a simple temporal filter. This makes it unusable for non-static camera datasets and results in artifacts in the backgrounds of the anonymization, as well as a halo effect when the masks are off. One way to address this is to utilize a deep in-painting pipeline [WZN*21] or a DGAN-based low-resolution deep in-painting pipeline [HH23] for filling in details behind the detected pedestrians, which can be further refined by taking into account the order of the generated pedestrians, in the way that DeepPrivacy2 utilizes it.

## 7 CONCLUSION

We propose an initial study in a pipeline for full-body anonymization of pedestrians in anomaly detection using diffusion-based generation that does not process the RGB images directly and thus it provides better privacy protection. Combining YoloV8 for object detection, ClipSeg and BiRefNet for segmentation, and OpenPose for pose estimation, the system produces context-aware, temporally consistent anonymizations that alter pedestrians' demographics while ensuring the diffusion model never processes non-anonymized data. We evaluated the impact on anomaly detection by training four models on data from our pipeline, DeepPrivacy2, and traditional methods like blurring and pixelization. Our approach achieves comparable results while fully anonymizing pedestrians. The approach still has drawbacks like sensitivity to incorrect pedestrian detection and tracking inputs, background blending artifacts, and overlapping. Future improvements will focus on minimizing artifacts, enhancing tracking, and refining skeletonization and color blending.

## 8 REFERENCES

[AJO24] ASRES M. W., JIAO L., OMLIN C. W.: Low-latency video anonymization for crowd anomaly detection: Privacy vs. performance, 2024. URL: https://arxiv.org/abs/2410.18717, arXiv:2410.18717.

[AZLL21] ASTRID M., ZAHEER M. Z., LEE J.-Y., LEE S.-I.: Learning not to reconstruct anomalies. *arXiv preprint arXiv:2110.09742* (2021).

[BDNM21] BIDSTRUP M., DUEHOLM J. V., NASROLLAHI K., MOESLUND T. B.: Privacy-aware anomaly detection using semantic segmentation. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II* (2021), Springer, pp. 110–123.

[BGU22] BUET-GOLFOUSE F., UTYAGULOV I.: Towards fair unsupervised learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), pp. 1399–1409.

[BKD*23] BIANCHI F., KALLURI P., DURMUS E., LADHAK F., CHENG M., NOZZA D., HASHIMOTO T., JURAFSKY D., ZOU J., CALISKAN A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023), pp. 1493–1504.

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 7291–7299.

[dONRM24] DE OLIVEIRA C. B., NEVES J. C., RIBEIRO R. O., MENOTTI D.: A multilevel strategy to improve people tracking in a real-world scenario. *arXiv preprint arXiv:2404.18876* (2024).

[FWZ*22] FAN J., WU K., ZHOU Y., ZHAO Z., HUANG S.: Fast model update for iot traffic anomaly detection with machine unlearning. *IEEE Internet of Things Journal 10*, 10 (2022), 8590–8602.

[HH23] HUANG L., HUANG Y.: Drgan: A dual resolution guided low-resolution image inpainting. *Knowledge-Based Systems 264* (2023), 110346.

[HL23a] HUKKELÅS H., LINDSETH F.: Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), pp. 1329–1338.

[HL23b] HUKKELÅS H., LINDSETH F.: Does image anonymization impact computer vision training? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 140–150.

[HML19] HUKKELÅS H., MESTER R., LINDSETH F.: Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing* (2019), Springer, pp. 565–578.

[HSML23] HUKKELÅS H., SMEBYE M., MESTER R., LINDSETH F.: Realistic full-body anonymization with surface-guided gans. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision* (2023), pp. 1430–1440.

[HZC*23] HE X., ZHU M., CHEN D., WANG N., GAO X.: Diff-privacy: Diffusion-based face privacy protection. *arXiv preprint arXiv:2309.05330* (2023).

[JCQ23] JOCHER G., CHAURASIA A., QIU J.: Ultralytics YOLO, Jan. 2023. URL: `https://github.com/ultralytics/ultralytics`.

[JZZ*23] JU X., ZENG A., ZHAO C., WANG J., ZHANG L., XU Q.: Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 15988–15998.

[KLY*21] KUANG Z., LIU H., YU J., TIAN A., WANG L., FAN J., BABAGUCHI N.: Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM international conference on multimedia* (2021), pp. 3182–3191.

[KMR*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W.-Y., ET AL.: Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 4015–4026.

[LCC*21] LV H., CHEN C., CUI Z., XU C., LI Y., YANG J.: Learning normal dynamics in videos with meta prototype network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15425–15434.

[LE22] LÜDDECKE T., ECKER A.: Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 7086–7096.

[LSJ13] LU C., SHI J., JIA J.: Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 2720–2727.

[LXW*20] LIU X., XIE L., WANG Y., ZOU J., XIONG J., YING Z., VASILAKOS A. V.: Privacy and security issues in deep learning: A survey. *IEEE Access 9* (2020), 4566–4593.

[LY24] LIN S., YANG X.: Animatediff-lightning: Cross-model diffusion distillation. *arXiv preprint arXiv:2403.12706* (2024).

[MELT20] MAXIMOV M., ELEZI I., LEAL-TAIXÉ L.: Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5447–5456.

[PBLM24] PIANO L., BASCI P., LAMBERTI F., MORRA L.: Latent diffusion models for attribute-preserving image anonymization. *arXiv preprint arXiv:2403.14790* (2024).

[PNH20] PARK H., NOH J., HAM B.: Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 14372–14381.

[RBL*21] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models, 2021. arXiv:2112.10752.

[RD16] RUCHAUD N., DUGELAY J.-L.: Automatic face anonymization in visual data: Are we really well protected? In *Electronic Imaging* (2016).

[RLR18] REN Z., LEE Y. J., RYOO M. S.: Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 620–636.

[SC21] STEED R., CALISKAN A.: Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021), pp. 701–713.

[SCS18] SULTANI W., CHEN C., SHAH M.: Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 6479–6488.

[SDABMP21] SUNDARARAMAN R., DE ALMEIDA BRAGA C., MARCHAND E., PETTRE J.: Tracking pedestrian heads in dense crowd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 3865–3875.

[SSS*24] SCHNEIDER D., SAJADMANESH S., SEHWAG V., SARFRAZ S., STIEFELHAGEN R., LYU L., SHARMA V.: Activity recognition on avatar-anonymized datasets with masked differential privacy, 2024. URL: https://arxiv.org/abs/2410.17098, arXiv:2410.17098.

[Ste] STEWART D.: Compel library. https://github.com/damian0815/compel. Accessed: 2024-06-15.

[SVH*24] STYPUŁKOWSKI M., VOUGIOUKAS K., HE S., ZIĘBA M., PETRIDIS S., PANTIC M.: Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), pp. 5091–5100.

[SZS24] SHI L., ZHANG J., SHAN S.: Anonymization prompt learning for facial privacy-preserving text-to-image generation. *arXiv preprint arXiv:2405.16895* (2024).

[vPPL*22] VON PLATEN P., PATIL S., LOZHKOV A., CUENCA P., LAMBERT N., RASUL K., DAVAADORJ M., NAIR D., PAUL S., BERMAN W., XU Y., LIU S., WOLF T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[VVdB17] VOIGT P., VON DEM BUSSCHE A.: The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing 10*, 3152676 (2017), 10–5555.

[WZH21] WANG D., ZHANG S., HUA G.: Robust pose estimation in crowded scenes with direct pose-level inference. *Advances in Neural Information Processing Systems 34* (2021), 6278–6289.

[WZN*21] WANG W., ZHANG J., NIU L., LING H., YANG X., ZHANG L.: Parallel multi-resolution fusion network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 14559–14568.

[XXH*21] XU R., XIAO Z., HUANG J., ZHANG Y., XIONG Z.: Edpn: Enhanced deep pyramid network for blurry image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 414–423.

[YSI*23] YUAN Y., SONG J., IQBAL U., VAHDAT A., KAUTZ J.: Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 16010–16021.

[ZGF*24] ZHENG P., GAO D., FAN D.-P., LIU L., LAAKSONEN J., OUYANG W., SEBE N.: Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407* (2024).

[ZLG*23] ZENG B., LIU X., GAO S., LIU B., LI H., LIU J., ZHANG B.: Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 628–637.

[ZLL*22] ZHAO M., LIU Y., LIU J., LI D., ZENG X.: Lgn-net: Local-global normality network for video anomaly detection. *arXiv preprint arXiv:2211.07454* (2022).

[ZOL*22] ZENG J., OUYANG H., LIU M., LENG L., FU X.: Multi-scale yolact for instance segmentation. *Journal of King Saud University-Computer and Information Sciences 34*, 10 (2022), 9419–9427.