DeepPrivacy2: Towards Realistic Full-Body Anonymization

Håkon Hukkelås   Frank Lindseth

Norwegian University of Science and Technology
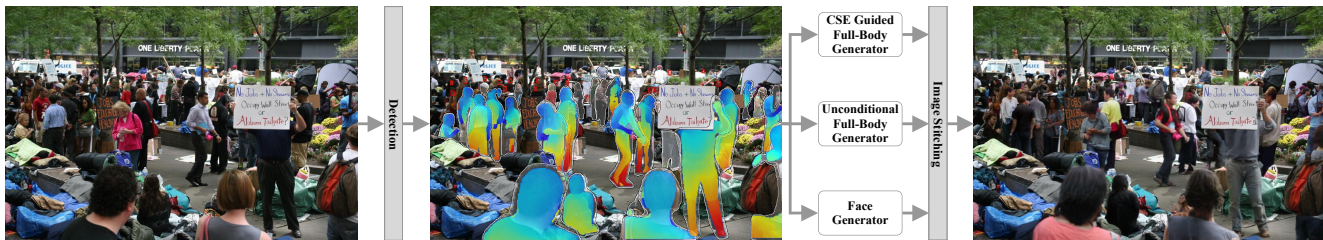
`hakon.hukkelas@ntnu.no`

Figure 1: DeepPrivacy2 detects and anonymizes individuals via three detection and synthesis networks; (1) a CSE-guided generator for individuals detected with dense pose (by CSE [30]), (2) an unconditional full-body generator for cases where CSE fails to detect (note the segmented persons without color-coded CSE detections), and (3) a face generator for the remaining individuals (marked in red). The original image is from Wider-Face [45].

Abstract

Generative Adversarial Networks (GANs) are widely adopted for anonymization of human figures. However, current state-of-the-art limits anonymization to the task of face anonymization. In this paper, we propose a novel anonymization framework (DeepPrivacy2) for realistic anonymization of human figures and faces. We introduce a new large and diverse dataset for full-body synthesis, which significantly improves image quality and diversity of generated images. Furthermore, we propose a style-based GAN that produces high-quality, diverse, and editable anonymizations. We demonstrate that our full-body anonymization framework provides stronger privacy guarantees than previously proposed methods. Source code and appendix is available at: github.com/hukkelas/deep_privacy2.

1. Introduction

Collecting and storing images is ubiquitous in our modern society, where a range of applications requires collecting privacy-sensitive data. However, collecting such data without anonymization or consent from the individual is troublesome due to recently introduced legislation in many areas (*e.g.* GDPR in EU). Traditional image anonymization (*e.g.* blurring) is widely adopted in practice; however, it severely distorts the data, making it unusable for future ap-

plications. Recently, *realistic anonymization* has been introduced as an alternative to traditional methods, where generative models can generate realistic faces fitting into a given context [7, 13, 26, 40]. However, current methods focus on face anonymization, which does not prevent recognition through identifiers outside the face, including both primary (*e.g.* ears, gait [15]) and secondary (*e.g.* gender) identifiers.

Surface Guided GANs (SG-GAN) [14] propose a full-body anonymization GAN guided on dense pixel-to-surface correspondences from Continuous Surface Embeddings (CSE) [30]. SG-GAN shows promising results for full-body anonymization, but their method often includes visual artifacts, degrading the image quality. The authors attribute the limited visual quality to the dataset, where they use a derivate of COCO [23] containing 40K human figures. Furthermore, the CSE segmentation used for anonymization does not include accessories/hair on the human body; thus, the anonymized individual often "wears" these unsegmented areas (see fig. 3). Additionally, SG-GAN fails to anonymize many individuals, as the CSE detector often fails to detect persons that are further away from the camera.

In this work, we extend Surface Guided GANs to address the limited visual quality and the insufficient anonymization due to poor segmentation. Furthermore, we address cases where the CSE detector fails to detect individuals. We summarize our contributions in the following.

First, we introduce the Flickr Diverse Humans (FDH) dataset. The FDH dataset consists of 1.5M images of human

Figure 2: Examples from the FDH dataset. Each image is annotated with keypoints, pixel-to-vertex correspondences (from CSE [30]) and a segmentation mask. The leftmost image shows annotations for the first image.

figures in diverse contexts extracted from the YFCC100M [42] dataset. We show that the larger dataset greatly benefits the visual quality of generated human figures.

Secondly, we propose a novel anonymization framework that combines detections from multiple modalities to improve the segmentation and detection of human figures. Our anonymization framework divides image anonymization into three individual anonymizers; (1) for human figures that are detected with a dense pose estimation (CSE), (2) for human figures that CSE does not detect, and (3) for the remaining faces (see fig. 1). For each category, our framework employs a simple inpainting GAN that follows established GAN training techniques for unconditional image generation [17, 18]. We show that our GAN generates high-quality and diverse identities with few task-specific modeling choices.

Finally, we extend our GAN for face anonymization on an updated version of the Flickr Diverse Faces (FDF) dataset [13]. In contrast to previous face anonymization techniques [7, 13, 26, 40], our GAN uses no pose guidance, enabling it to anonymize individuals where pose information is challenging to detect. Furthermore, we show that our style-based generator can adapt methods from unconditional GANs to find global semantically meaningful directions in the GAN latent space. This enables text-guided attribute editing for our anonymization pipeline.

DeepPrivacy2 surpasses all previous state-of-the-art realistic anonymization methods in terms of image quality and anonymization guarantees. We validate the effectiveness of DeepPrivacy2 with extensive qualitative and quantitative evaluation. Our code, pre-trained models, and the FDH dataset is available at github.com/hukkelas/deep_privacy2.

2. Related Work

Image Anonymization   Naive image anonymization (*e.g.* masking, blurring, pixelation) is widely adopted in practice; however these methods severely degrade the quality of the anonymized image, making the data unusable for many applications. Early work focused on the K-same family of algorithms [8, 16, 31], which provides better pri-

vacy guarantees and data usability than naive methods , but generate highly corrupted images. Recent work on deep generative models reflects that learning-based anonymization can realistically anonymize data while retaining its usability for downstream applications. These methods anonymize face regions by either inpainting missing regions [13, 26, 40, 41] or transforming [7, 35] the original face. Our method anonymizes by inpainting, as inpainting-based methods provide stronger privacy guarantees than transformative methods, as they never observe the original privacy-sensitive information. The majority of prior work focuses on face anonymization, which compromises privacy for many use cases, as they leave several primary (*e.g.* ears, gait) and secondary (*e.g.* gender) identifiers on the human body untouched. There is a limited amount of work focusing on full-body anonymization [2, 14, 26], where prior methods are limited to low-resolution images [26] or generate images with visual artifacts [2, 14].

Full-body Synthesis   Recent work on full-body synthesis focus on limited tasks, such as transferring source appearances into new poses [1, 3, 22, 29, 38], with different garments [9, 37, 38], or with new motion [3]. These methods are often guided on dense pixel-to-surface correspondences or sparse keypoints annotations. In contrast to these methods, our anonymization approach does not rely on a source appearance to transfer, and the majority of the aforementioned methods do not handle large variations in background contexts. Furthermore, a number of these methods focus on low-variance datasets (*e.g.* DeepFashion [24]), which consists of a limited number of identities in similar poses and a close-to static context (white background). There is a limited amount of work focusing on full-body synthesis without a source appearance, where Ma *et al.* [25] proposes a pose-guided GAN for novel full-body synthesis.

3. The Flickr Diverse Humans Dataset

The Flickr Diverse Humans (FDH) dataset consists of 1.53M images of human figures from the YFCC100M [42] dataset. Figure 2 shows examples from the dataset. Each image contains a single human figure in the center, with

a pixel-wise dense pose estimation from CSE [30], 17 keypoint annotations from a keypoint R-CNN model [10], and a segmentation mask. The segmentation mask is the union of the mask from a CSE detector and Mask R-CNN [10] trained on COCO. The dataset is automatically filtered through confidence thresholding, automatic image quality assessment, the number of body parts visible in the image, and overlap between keypoint and CSE predictions (see Appendix A for more details). Otherwise, we perform no further filtering such that the dataset includes individuals in all various contexts. The resolution of each image is 288×160, and the dataset is split into 1,524,845 images for training and 30K images for validation. Compared to previously adopted datasets for full-body synthesis [14, 24, 47], FDH is much larger and contains a diverse set of individuals from in-the-wild images. Additionally, FDH is less curated than typical datasets for generative modeling, where it includes human figures with unusual poses, perspectives, lighting conditions, and contexts. This is to ensure that our anonymization method can handle such conditions.

4. The DeepPrivacy2 Anonymization Pipeline

This section outlines the core technologies used for our anonymization pipeline. First, we present our ensemble detection pipeline, then our GAN-based synthesis method.

4.1. Detection

The main objective of the detection module is to ensure that all individuals in the image are detected. DeepPrivacy2 uses an ensemble of three detection networks from different modalities; DSFD [21] for face detection, CSE [30] for dense pose estimation, and Mask R-CNN [10] for instance segmentation. The pipeline categorizes the detections into three categories; individuals with dense pose estimation (detection w/ CSE), individuals not detected by CSE (detection w/o CSE), and faces that are not included in the former categories. For each category, we propose individual anonymization methods, introduced in section 4.2. For human figures, we anonymize the union of Mask R-CNN and CSE segmentations, such that accessories/hair detected by Mask R-CNN (but not CSE) are anonymized (see fig. 3). Note that dense pose estimation is not essential for privacy, but it substantially improves synthesized image quality. Furthermore, the detections are tracked with a bounding box tracker, such that the anonymization can retain the same identity over a sequence of frames. Compared to SG-GAN [14], the ensemble of detectors significantly improves detection recall, as DeepPrivacy2 uses Mask R-CNN and DSFD for fallback detection when the CSE detector fails.

Implementation Details. Instance segmentation and CSE segmentation are combined via simple Intersection over Union (IoU) thresholding, where we assume all detections with an IoU higher than 0.4 are the same individual.

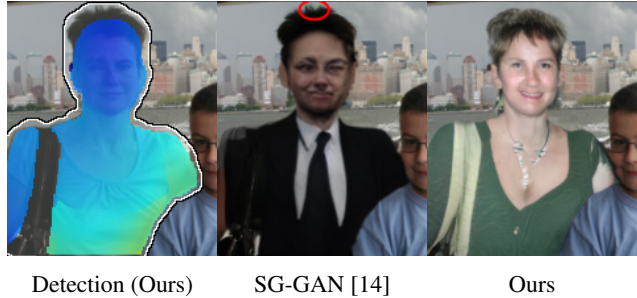Detection (Ours)   SG-GAN [14]   Ours

Figure 3: SG-GAN [14] anonymizes only the area from a CSE segmentation (marked in blue tint), which does not include accessories/hair. This results in SG-GAN [14] anonymization often wearing the original hair/accessories of the original identity (marked in red). In contrast, DeepPrivacy2 anonymizes the segmentation from Mask R-CNN (outlined), which includes hair and clothing.

All instance segmentations from Mask R-CNN not combined with a CSE detection are categorized as a detection without CSE. All face detections within the CSE or Mask R-CNN segmentation are discarded. All detections are tracked with simple Kalman filtering on bounding boxes, following the implementation of motpy [28]. We use Mask R-CNN and the CSE implementations from detectron2 [43], specifically, the ResNeXt-101 FPN [44] Mask R-CNN, and ResNet-101 [11] CSE. We adapt DSFD [21] from the official implementation of the authors.

4.2. Synthesis Method

DeepPrivacy2 uses three independently trained generators for the three different detection categories introduced in section 4.1 (detection w/ cse, detection w/o cse, faces). While the tasks significantly differ in complexity, they share training setup and architecture to a high degree. Here, we first present our style-based generator, then present task-specific modeling choices for full-body and face synthesis. Each generator frame the anonymization task as an image inpainting task, where we remove areas to be anonymized and let a generator complete the missing region. Specifically, the input and output of each generator is given by [1],

$$\tilde{I} = G(I \odot M, M, z) \odot M + (1 - M) \odot I, \quad (1)$$

$\odot$ is element-wise multiplication, I the original image, and M indicates missing regions (M is 1 for known pixels and 0 for pixels to be anonymized).

---

[1]For the CSE-guided generator, the CSE-embedding is concatenated with $I \odot M$ to the input of the generator. See section 4.2.1.

| (a) Detections | (b) Descending ordering | (c) Ascending ordering |

Figure 4: Anonymization results comparing our method with descending and ascending image stitching order. The ascending ordering stitches foreground objects last, which improves image quality at detection borders (*e.g.* marked in red).

4.2.1 A Style-Based U-Net Generator

Our synthesis method follows the implementation of Surface-guided GANs [14]. The generator is a U-Net [36] with limited task-specific modeling choices, consisting of a context encoder and a style-based decoder. The context encoder uses a sequence of convolutions and downsampling layers, with residual connections at every feature map resolution. We use no normalization layers in the encoder, as it performs similarly without it. However, we find it essential to apply instance normalization for the features in the U-net skip connections, where we combine features from the encoder and decoder as additive residuals. The decoder follows the design of Stylegan2 [18], with the operation order *instance normalization → convolution → style modulation*. Note that we replace the baked-in weight demodulation in Stylegan2 with instance normalization, as we find that normalization on expected statistics works poorly when large areas of the input are missing.

Furthermore, we increase the depth of the U-net to 5 downsampling layers, such that the minimum feature resolution is 9×5. In contrast, SG-GAN has three downsampling blocks, where most parameters are at the 36×20 resolution. We observe no performance degradation by increasing the depth while improving inference speed, as more parameters are located at lower resolution layers. Finally, we remove V-SAM and discriminator surface supervision used in SG-GAN. Appendix C includes further details.

Full-Body Synthesis The full-body generator is trained on the FDH dataset at a resolution of 288×160. We train two independent generators for full-body synthesis; one that concatenates the CSE embedding to the input image and one that does not. The CSE embedding has a resolution of $16 \times 288 \times 160$, where we use the pixel-to-vertex embedding map released in the official implementation of CSE [30, 43].

Face Synthesis In contrast to previous face anonymization methods [13, 26, 40], we propose a generator that does not use keypoints for synthesis. Removing the keypoint detector improves detection recall in cases where keypoints are difficult to detect. We train the face generator on an updated version of the FDF dataset [13], which increases the image resolution to 256×256 from the original 128×128.

4.3. Recursive Stitching

The final stage of our pipeline is pasting the anonymized identities into the original image. Unlike face anonymization, full-body anonymization has many detection overlaps. If not handled correctly, these overlaps generates visually annoying artifacts at the border between individuals.

Our stitching approach recursively stitches each individual in ascending order depending on the number of pixels the person covers. The recursive stitching assumes that the synthesis method handles overlapping artifacts when generating each individual. Additionally, our ordering assumes that objects in the foreground cover a larger area, where foreground objects are stitched in last. The reverse order (foreground objects first) results in background objects "overwriting" foreground objects, as the detections can overlap (see fig. 4). This naive ordering significantly reduces visual artifacts at borders between individuals.

5. Experimental Evaluation

We validate our proposed anonymization pipeline in terms of synthesis quality, using anonymized data for future development, and anonymization guarantees. There are no standard baselines to compare against for realistic anonymization of data. Thus, we compare against traditional anonymization techniques, and DeepPrivacy [13], a widely adopted realistic face anonymizer. Additionally, we compare our full-body generator to Surface Guided GANs [14]. Appendix D includes random anonymized images on Cityscapes [4], COCO [23], FDH, and FDF256.

Experimental Details. All models are trained with Pytorch 1.10 [33] on 4 NVIDIA V100-32GB. For qualitative

(a) Detected (b) No CSE (c) (d) (e)

Figure 5: Synthesis results on FDH. (a) is the original identity and the anonymization mask, (b) is the unconditional generator, and (c-e) is the CSE-guided generator

examples, we use multi-modal truncation to improve image quality while retaining diversity [27]. We report Fréchet Inception Distance (FID) [12] and FID_{CLIP} [2] [20] to evaluate image quality using Torch Fidelity [32]. The three generators (for CSE-guided, unconditional and face) has 43M parameters each.

Datasets. For training, we use the FDH dataset (see section 3) for full-body synthesis, and FDF256 for face synthesis. The FDF256 dataset is an updated version of FDF [13],

---

[2]FID_{CLIP} is less sensitive to ImageNet classes. ImageNet-FID is insensitive to faces and scores images containing ImageNet objects (*e.g.* tie) higher [20].

Figure 6: Synthesis results on FDF256. First column shows the original identity and the anonymization mask. Columns 2-5 shows generated identities from DeepPrivacy2.

where the image resolution is increased from 128×128 to 256×256 (see Appendix B). For evaluation, we use Market1501 [47], Cityscapes [4], and COCO [23]. We follow the standard train/validation split for all datasets.

Runtime Analysis. The DeepPrivacy2 architecture is computationally efficient, where the CSE-guided generator processes ~ 11.6 frames per second (FPS), and the face generator at ~ 7.9 FPS on an NVIDIA 1080 8GB GPU. In contrast, the SG-GAN [14] generator processes ~ 7.3 FPS, where our improved runtime originate from the removal of V-SAM and moving the majority of parameters to lower resolution layers. The entire pipeline (detection, synthesis and stitching) require ~ 2.8 seconds to process an image with 12 persons on an NVIDIA 1080 8GB GPU.

5.1. Synthesis Quality

Full-Body Synthesis. Figure 5 shows diverse synthesized examples on the FDH dataset. Our model generates high-quality human figures that seamlessly transition into the original image. Furthermore, the model can handle a large variety of background contexts, poses, and overlapping objects. We find CSE guidance necessary for high-quality anonymization, where the unconditional generator often generates human figures with unnatural poses (see fig. 5). This is reflected in quantitative metrics, where
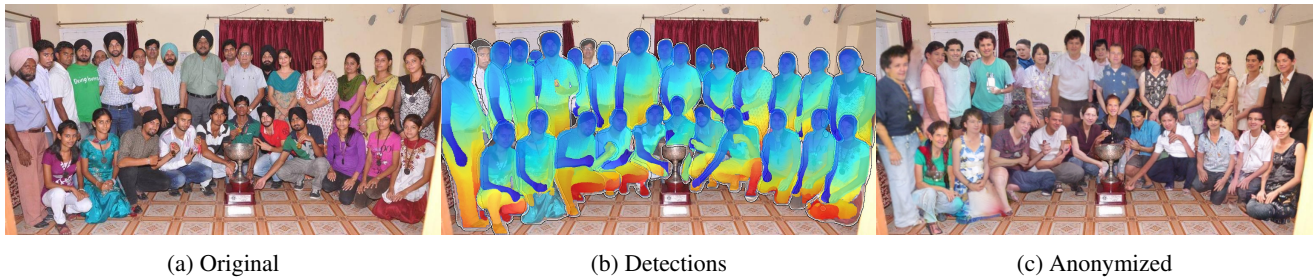
1333

| (a) Original | (b) Detections | (c) Anonymized |

Figure 7: DeepPrivacy2 anonymization result on an image from WIDER-Face. Appendix D includes random examples.

| (a) Original | (b) DeepPrivacy | (c) Ours |

Figure 8: Face anonymization comparison between our method and DeepPrivacy [13].
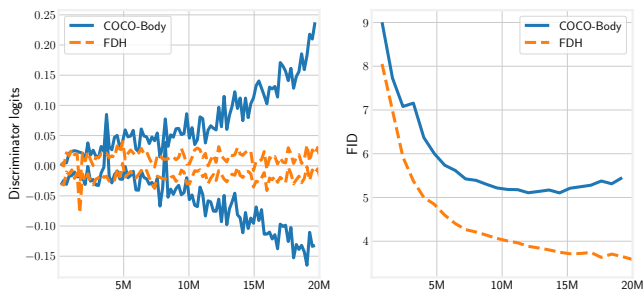

Figure 9: Our generator overfits early when trained on the small COCO-Body dataset (blue) [14], which consists of $\sim 40K$ images of human figures. Note that this occurs even with the strong data augmentation used by [14]. No data augmentation, other than horizontal flip, is used for FDH.

the CSE-guided generator (FID=5.6, FID$_{\text{CLIP}}$=1.7) substantially improves over the unconditional generator (FID=6.1, FID$_{\text{CLIP}}$=2.30). Furthermore, the primary improvement of our model compared to Surface-guided GANs [14] is the larger and more diverse FDH dataset. The same model trained on the COCO-Body dataset [14] starts to overfit early in training, reflected by the diverging discriminator logits and increasing FID (fig. 9).

Face Synthesis. Figure 6 shows generated results on the FDF256 dataset. Directly comparing our face anonymizer to DeepPrivacy [13] is not straightforward, as we synthe-

size higher resolution images (256×256, not 128×128). Additionally, the FDF256 dataset does not include the same images as the original dataset, as FDF256 filter out lower resolution images. Nevertheless, to validate our modeling choices, we retrain our GAN for 128×128 synthesis on FDF [13]. Our GAN achieves a FID of 0.56, a significant improvement compared to DeepPrivacy (FID=0.68) [13]. Note that this is without using face keypoints, which the original DeepPrivacy uses to improve quality.

Figure 8 compares the open-source DeepPrivacy [13] to our method. Our method generates higher quality faces and handles overlaps between detections better. Also, note that DeepPrivacy does not anonymize all faces in the image, as it is unable to detect keypoints for all individuals [3].

Attribute-Guided Anonymization DeepPrivacy2 allows for controllable anonymization through text prompts by adapting StyleMC [19]. StyleMC finds global semantically meaningful directions in the GAN latent space by manipulating images towards a given text prompt with a CLIP-based [34] loss. Figure 10 shows attribute-guided anonymization, where the global directions are found over 256 images. As far as we know, DeepPrivacy2 is the first to enable controllable anonymization through text prompts, whereas previous methods are limited to no control or attribute preservation from the original identity [7] .

---

[3]Even with confidence threshold of 0.05, DeepPrivacy is unable to detect keypoints for all individuals.

Original Anonymization | Happy

Mustache | Blue eyes

Figure 10: Latent manipulations with StyleMC [19]. The text prompt used for each edit is shown below each image. See our video for an interactive demo: https://youtu.be/faoNyaaORts.

Anonymization	R1 ↓	mAP ↓
Original	94.4	82.5
Pixelation 8×8	54.6	16.1
Pixelation 16×16	70.3	36.6
Mask-out	45.5	**8.0**
SG-GAN [14]	74.4	30.2
Full-body anonymization (Ours)	**44.7**	8.5

Table 1: Re-identification mAP and rank-1 accuracy on Market1501 [47] using OSNet [48].

5.2. Anonymization Evaluation

Anonymization Guarantee To evaluate the anonymization guarantee of DeepPrivacy2, we evaluate how well automatic re-identification tools can identify anonymized individuals. Specifically, we evaluate the re-identification rate of OSNet [48] by anonymizing Market1501 [47]. In this case, a lower re-identification mAP and rank-1 accuracy (R1) reflects worse re-identification, indicating improved anonymization. Appendix C details the experiment further.

Dataset	Train w/ Anon. Data		Validate w/ Anon. Data	
	Box AP ↑	Kp. AP ↑	Box AP ↑	Kp. AP ↑
Original	53.6	64.0	53.6	64.0
Masked Out	10.1	0.5	17.0	1.8
Pixelation 8×8	10.4	1.0	29.1	2.2
Pixelation 16×16	10.1	1.5	36.5	12.0
DeepPrivacy2 (w/o CSE)	21.4	10.2	**49.9**	11.5
DeepPrivacy2	**26.0**	**31.9**	49.4	**48.4**

Table 2: Keypoint (Kp.) AP on the COCO [23] validation set with a Keypoint R-50 FPN R-CNN [10].

Table 1 reflects that pixelation enables re-identification of several of the anonymized individuals. Our full-body anonymizer yields similar anonymization guarantees as masking out the area and significantly improve compared to pixelation. Furthermore, the full-body anonymization of SG-GAN [14] provides poorer anonymization results than DeepPrivacy2. This is caused by the CSE detector failing in several cases and the poor segmentation of accessories/hair in SG-GAN, where the anonymized identity often "wears" parts of the original identity (see section 4.1).

5.2.1 Training and Evaluating on Anonymized Data

A typical use case for anonymization is collecting and anonymizing data for the development of computer vision systems. We evaluate DeepPrivacy2 on two established computer vision benchmarks: COCO [23] person keypoint estimation and Cityscapes [4] instance segmentation. We evaluate two use cases; (1) using anonymized data for training, and (2) using anonymized data for validation with a pre-trained model. For the former, we report evaluation metrics on the original validation dataset.

COCO Person Keypoint Estimation. Table 2 analyzes the effect of anonymization on the COCO dataset for person keypoint estimation. Pixelation greatly affects model training for the fine-grained task of keypoint estimation, whereas DeepPrivacy2 significantly improves over traditional methods. Note that CSE fails to detect many individuals in the COCO dataset, which yields poor pose preservation for individuals anonymized by the unconditional generator.

Cityscapes Instance Segmentation. Table 3 analyzes the effect of anonymization on the Cityscapes dataset. DeepPrivacy2 improves over pixelation and mask-out, but the gap is less prevalent than for keypoint estimation. We believe this originates from model weight initialization [4].

Is surface guidance necessary? Section 5.1 established that the CSE-guided generator improves image quality compared to the unconditional generator. We now ask the question; does the improved image quality translate to improvements when using the anonymized data? In table 3 and ta-

---

[4]The Cityscapes model is initialized from a COCO pre-trained Mask R-CNN, while the keypoint R-CNN from an ImageNet [5] backbone.

	Train w/ Anon. Data		Validate w/ Anon. Data	
Dataset	mAP ↑	AP$_{person}$ ↑	mAP ↑	AP$_{person}$ ↑
Original	36.5	35.0	36.5	35.0
Masked Out	34.0	26.4	27.7	4.7
Pixelation 8×8	34.7	27.1	29.4	10.2
Pixelation 16×16	34.7	29.6	32.0	21.8
DeepPrivacy2 (w/o CSE)	33.4	27.5	33.1	**27.8**
DeepPrivacy2	**35.2**	**30.3**	**33.2**	27.3

Table 3: Instance segmentation AP on the Cityscapes [4] validation set with a Mask R-CNN [10] R-50 FPN.

ble 2, we replace the CSE-guided generator with the unconditional generator, such that all persons are anonymized without CSE-guidance (denoted *DeepPrivacy2 w/o CSE*). Removing CSE-guidance severely hurts performance, especially when using the anonymized data for training.

6. Conclusion

DeepPrivacy2 is an automatic realistic anonymization framework for human figures and faces, and is a practical tool for anonymization without degrading the image quality. Compared to previously proposed anonymization frameworks, we show that DeepPrivacy2 substantially improves image quality and privacy guarantees. Furthermore, we introduce the FDH dataset, a large-scale full-body synthesis dataset that includes a wide variety of identities in different poses and contexts. Our new FDH dataset, combined with our simple style-based GAN, improves image quality and diversity of human figure synthesis for in-the-wild images. Furthermore, we show that our simple style-based GAN generates high-quality human faces that are controllable through user-guided anonymization via text prompts. We believe that our open-source framework will be a useful tool for computer vision researchers and other entities requiring anonymization while retaining image quality.

Societal Impact Recently introduced legislation in many regions has complicated collecting privacy-sensitive data, where consent from individuals is required for storing the data. This can act as a barrier for developing applications relying on high-quality images, such as computer vision models. This paper proposes an automatic realistic image anonymization framework that simplifies the collection of privacy-sensitive data while retaining the original image quality. We believe this will be a highly useful tool for the computer vision field. Nevertheless, our work focuses on synthesizing realistic humans, which has a potential for misuse (*e.g.* DeepFakes). There is a large focus in the community to mitigate this, for example, the DeepFake Detection Challenge [6] and model watermarking [46].

Figure 11: The generator samples from a small subset of different identities given the condition.

6.1. Limitations

DeepPrivacy2 generates a limited set of identities given a particular input condition. The input condition is highly descriptive of the shape of the original identity and the context that the identity should fit into. Thus, the generator learns a sampling probability of identities given the condition. For example, if the generator observes a baseball field, the synthesized identity is likely to be a baseball player (fig. 11).

As with any anonymization framework, DeepPrivacy2 cannot guarantee anonymization without human supervision, as the detector can fail. However, DeepPrivacy2 uses a set of detectors from different modalities to improve detection in cases where one or more of the detectors fail. Also, DeepPrivacy2 uses dense pose description for anonymization, which allows identity recognition through gait [15].

Synthesis Quality DeepPrivacy2 significantly improves full-body synthesis for in-the-wild images; however, it struggles in several scenarios. First, DeepPrivacy2 relies on dense pose estimation to synthesize high-quality human figures, where the image quality is severely degraded in cases where the pose description is incorrect. Furthermore, we find our full-body GAN harder to edit (*e.g.* attribute edit via text prompts [19]), and we observe that common directions in the latent space do not translate to semantically equivalent transformations for different poses/background contexts.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing Images of Humans in Unseen Poses. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8340–8348, 2018.

[2] 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.

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. Everybody Dance Now. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2019.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[5] 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 (CVPR)*, pages 248–255, 2009.

[6] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. 2020.

[7] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live Face De-Identification in Video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9377–9386, 2019.

[8] Ralph Gross, Latanya Sweeney, F. de la Torre, and Simon Baker. Model-Based Face De-Identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 161–161, 2006.

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE, oct 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[14] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. Realistic full-body anonymization with surface-guided gans. *arXiv preprint arXiv:2201.02193*, 2022.

[15] Arun Jain, Anil and Flynn, Patrick and Ross. *Handbook of Biometrics*. Springer US, Boston, MA, 2008.

[16] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. *2015 International Conference on Biometrics (ICB)*, pages 278–285, 2015.

[17] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[19] Umut Kocasari, Alara Dirik, Mert Tiftikci, and Pinar Yanardag. StyleMC: Multi-Channel Based Fast Text-Guided Image Generation and Manipulation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3441–3450, 2022.

[20] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechet Inception Distance. *arXiv preprint arXiv:2203.06026*, 2022.

[21] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: Dual Shot Face Detector. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064, 2019.

[22] Yining Li, Chen Huang, and Chen Change Loy. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3688–3697, 2019.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[24] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[25] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled Person Image Generation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 99–108, 2018.

[26] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. CIA-GAN: Conditional Identity Anonymization Generative Ad-

versarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455, 2020.

[27] Ron Mokady, Michal Yarom, Omer Tov, Oran Lang, Daniel Cohen-Or, Tali Dekel, Michal Irani, and Inbar Mosseri. Self-Distilled StyleGAN: Towards Generation from Internet Photos. 2022.

[28] Wiktor Muron. motpy - simple multi object tracking library. https://github.com/wmuron/motpy, 2022.

[29] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos. Dense Pose Transfer. In *Computer Vision – ECCV 2018*, pages 128–143. Springer International Publishing, 2018.

[30] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous Surface Embeddings. In *Advances in Neural Information Processing Systems*, pages 17258—-17270. Curran Associates, Inc., 2020.

[31] E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[32] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch. https://github.com/toshas/torch-fidelity, 2020.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748—-8763, 2021.

[35] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *Computer Vision – ECCV 2018*, pages 639–655. Springer International Publishing, Cham, 2018.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[37] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and Pose Control for Image Synthesis of Humans from a Single Monocular View. *arXiv preprint arXiv:2102.11263*, feb 2021.

[38] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural Re-rendering of Humans from a Single Image. In *Computer Vision – ECCV 2020*, pages 596–613. Springer International Publishing, Cham, 2020.

[39] Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure, 2019.

[40] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5050–5059, 2018.

[41] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A Hybrid Model for Identity Obfuscation by Face Replacement. In *Computer Vision – ECCV 2018*, pages 570–586. Springer International Publishing, Cham, 2018.

[42] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, jan 2016.

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[44] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017.

[45] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. WIDER FACE: A Face Detection Benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[46] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14428–14437, 2021.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[48] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3701–3711, 2019.