RAD: Realistic Anonymization of Images Using Stable Diffusion

Simon Malm*
Linköping University
Linköping, Sweden

Viktor Rönnbäck*
Linköping University
Linköping, Sweden

Amanda Håkansson
Axis Communications
Linköping, Sweden

Minh-Ha Le
Linköping University
Linköping, Sweden

Karol Wojtulewicz
Linköping University
Linköping, Sweden

Niklas Carlsson†
Linköping University
Linköping, Sweden

Abstract

Many of the deep learning models currently driving the advancements in computer vision, expected to transform our society, require extensive training data. However, privacy regulations require explicit consent or anonymization of personal data, and traditional anonymization methods degrade data quality, thus hindering model performance. To address this challenge, we introduce the Realistic Anonymization using Diffusion (RAD) framework, which uses Stable Diffusion and ControlNet to produce high-quality synthetic images. RAD's three-step pipeline maintains contextual integrity and data utility, achieving superior image quality compared to previous GAN-based methods. We evaluated RAD's privacy preservation and data utility through face recognition accuracy, a segmentation task, and human assessment. RAD anonymized faces in 95.5% of cases, with high photo-realism ratings from human evaluators. Segmentation tasks on both original and anonymized images showed minimal performance drop, confirming RAD's high utility. Our analysis also identifies the strengths and weaknesses of using Stable Diffusion for full-body anonymization in various conditions. In summary, our work advances the understanding of high-utility anonymized data generation, and demonstrates that RAD can effectively balance privacy and utility.

CCS Concepts

• **Security and privacy**; • **Applied computing**; • **Computing methodologies** → **Computer vision**;

Keywords

Anonymization, Realistic Images, Stable Diffusion

ACM Reference Format:
Simon Malm, Viktor Rönnbäck, Amanda Håkansson, Minh-Ha Le, Karol Wojtulewicz, and Niklas Carlsson. 2024. RAD: Realistic Anonymization of Images Using Stable Diffusion. In *Proceedings of the 23rd Workshop on Privacy in the Electronic Society (WPES '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3689943.3695048

*Both authors contributed equally to this research.
†Corresponding author: Niklas Carlsson, niklas.carlsson@liu.se

WPES '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-1239-5/24/10
https://doi.org/10.1145/3689943.3695048

1 Introduction

Recent advancements in deep learning have revolutionized computer vision, enabling sophisticated applications across industries like security, transportation, manufacturing, and healthcare. However, despite the promising potential of these models, their effectiveness is limited by the lack of realistic anonymization techniques.

To see this, we note that these models typically rely on large, diverse datasets, which are difficult to obtain due to strict privacy regulations like the General Data Protection Regulation (GDPR) [27]. With GDPR mandating anonymization or explicit consent for datasets containing identifiable individuals, and traditional anonymization techniques like blurring or pixelization only providing privacy protection by downgrading the data quality, there is a need for more sophisticated methods that can achieve high-utility anonymization.

Goal: The central challenge in the anonymization of datasets is to achieve a desirable balance between privacy and utility. Ideally, we want realistic anonymization that replaces real individuals in images with highly realistic but generated ones, preserving data utility while also proving strong protection of the individual from identification (i.e., achieving a high privacy level). By maintaining visual coherence, such anonymization methods have the potential to create high-quality training data that complies with privacy regulations without sacrificing performance.

Main Contributions: As a step in this direction, we present the design and implementation of the Realistic Anonymization using Diffusion (RAD) framework (Section 4) and use a diverse set of experiments (Section 5) to demonstrate its effectiveness and provide insights into several important tradeoffs associated with RAD, including everything from individual components of the system to the privacy-utility tradeoff of the generated images themselves.

Design & Implementation: RAD employs a three-step pipeline (Section 4) to preserve contextual integrity and data utility. The process starts with pre-processing, where we detect people and extract structural features. In the synthesis step, Stable Diffusion, conditioned by the extracted ControlNet features, generates a synthetic image. Finally, the synthesized image is seamlessly integrated with the original background during the stitching step, ensuring realistic anonymization without compromising the original scene. By combining Stable Diffusion with a method to cut out and stitch back the generated person, RAD produces high-quality images of higher resolution than presented by most prior works on full-body anonymization, most of which use GANs and have considered lower image resolutions than us (Section 7). In comparison, RAD is designed for larger image sizes (e.g., >512×512 pixels) and excels at

generating people that are larger in the frame. The source code is shared with the paper: https://github.com/viktorronnback/RAD.

Effective Anonymization: To evaluate RAD's anonymization effectiveness, we used two methods (Section 5.3) to measure the achieved privacy level: face recognition accuracy and human evaluation. For face recognition, the FaceNet512 model measured the cosine distance between original and anonymized faces, showing successful anonymization in 95.5% of cases. A survey with 67 participants provided human evaluation. Participants first identified the most similar anonymized face from a set, showing that higher anonymization strength reduced correct identifications. They then rated anonymization effectiveness, which correlated with face embedding distances, confirming RAD's ability to preserve privacy while retaining data utility.

High Image Utility: To evaluate the utility of the generated images, we employed two methods (Section 5.4): training instance segmentation models with anonymized images and assessing their photo-realism through human evaluation. First, two YOLOv8 segmentation models were trained on the Cityscapes dataset: one with original images and the other with anonymized images. The comparison showed a slight decrease in detection and segmentation accuracy for the anonymized data, particularly for the person class, but most of the data's utility was preserved, evidenced by the minor drop in performance. Human evaluation rated the photo-realism of anonymized images with an average score of 3.96 out of 5, suggesting that the images retain sufficient quality for training data in computer vision tasks. Notably, there was no correlation between photo-realism and anonymization scores, indicating that higher privacy levels did not compromise perceived realism. In summary, while our method does not offer formal privacy guarantees, our combined privacy and utility evaluation results demonstrate the feasibility and realism of the anonymization pipeline.

Image Generation Insights: We also provide insights into the strengths and weaknesses of using Stable Diffusion models for the task (Section 5.2). For example, the strength of Stable Diffusion was shown to lie in its capability for full-body anonymization across diverse poses, backgrounds, and lighting conditions, effectively modifying features like hairstyle, facial features, and clothing while often retaining gender and ethnicity cues. In datasets like Pexel-Humans, which feature close-up, identifiable subjects, anonymization significantly alters personal details, such as tattoos or clothing. However, the synthesis process can introduce artifacts such as extra limbs or lighting issues, impacting image utility, albeit fixable. Segmentation and detection failures in preprocessing can also lead to incomplete anonymization. In datasets like Cityscapes, with smaller, distant targets, challenges include generating detailed faces and accurately discerning poses. Despite these hurdles, RAD's ability to preserve background context while anonymizing individuals makes it practical for annotated images, though maintaining identifiable body shapes poses potential privacy risks in certain contexts.

Outline: After a brief background (Section 2), we present RAD and its three-step anonymization pipeline (Section 4). We then present our performance evaluation (Section 5), spanning selected insights into the strengths and weaknesses (Section 5.2), the achieved privacy (Section 5.3), and the achieved utility (Section 5.4). Finally, we provide a discussion (Section 6) and put the work in the context of related works (Section 7), before concluding the paper (Section 8).

2 Background

2.1 Anonymization

Traditional image anonymization methods like blurring, masking, or pixelization have been widely used for privacy protection, albeit with mixed effectiveness [23]. However, these techniques significantly degrade image quality, which can hinder their utility in subsequent tasks such as training deep learning models [15]. Examples in Figure 1 illustrate these limitations.

Realistic Anonymization: Realistic anonymization, sometimes also called feature-preserving anonymization and generative de-identification, aims to replace individuals in images with realistic synthetic counterparts [14], thereby maintaining the images' utility for downstream tasks.

Privacy-Utility Tradeoff: The privacy-utility tradeoff refers to the balance between data anonymity and its usefulness when applying data manipulation algorithms [25]. This concept is extensively studied in fields like data mining and healthcare, where preserving privacy while maintaining data utility is critical [28]. In the context of using images for training data, the goal is to achieve anonymity comparable to methods like masking out (Figure 1(c)), while retaining the utility of the original image (Figure 1(a)). Realistic anonymization techniques aim to achieve this balance, though quantifying it remains challenging due to the absence of universal metrics for privacy and utility assessment, which often vary based on specific applications. In this paper, we use both objectively evaluated automatic tasks (facial recognition and segmentation) and more subjective evaluation by humans to evaluate our system and its privacy-utility tradeoff.

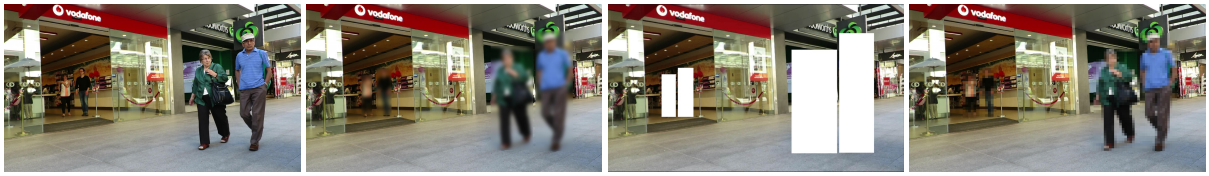
2.2 Stable Diffusion and ControlNet

Latent Diffusion Models: Diffusion models, especially latent diffusion models, have become increasingly popular recently. By compressing data into a lower-dimensional latent space, they focus on semantically relevant aspects of images, significantly reducing computational complexity [9, 31].

Stable Diffusion: Stable Diffusion models, a subset of latent diffusion models, feature specific architectural enhancements. Introduced publicly in August 2022 by Stability AI, Runway ML, and CompVis [1], Stable Diffusion v1 has since been enhanced with numerous improved models and architectures developed by both the core team and the community.

Stable Diffusion XL (SDXL), proposed by Podell et al. [29], represents an enhanced model and architecture within the Stable Diffusion framework. It incorporates additional conditioning of the U-Net model during training, considering image size, crop dimensions, and aspect ratios. Unlike previous models limited to 512×512 pixels, SDXL is trained on images up to 1024×1024 pixels, thereby producing higher-quality outputs for larger image sizes.

ControlNet: ControlNet, proposed by Zhang et al. [36], enhances user control over diffusion processes beyond traditional text prompts and basic parameters. Unlike existing methods, ControlNet utilizes various image types like edge maps, human pose skeletons, depth maps, or normal maps for conditioning. It achieves this by duplicating diffusion blocks, updating the weights of these copies independently while keeping the original diffusion block unchanged ("locked"). Outputs from these paired blocks are then



(a) Original (b) Blurring (c) Masking out (d) Pixelization

Figure 1: Three common anonymization techniques: (b) blurring, (c) masking out, and (d) pixelization. Image from MOT17 [26].

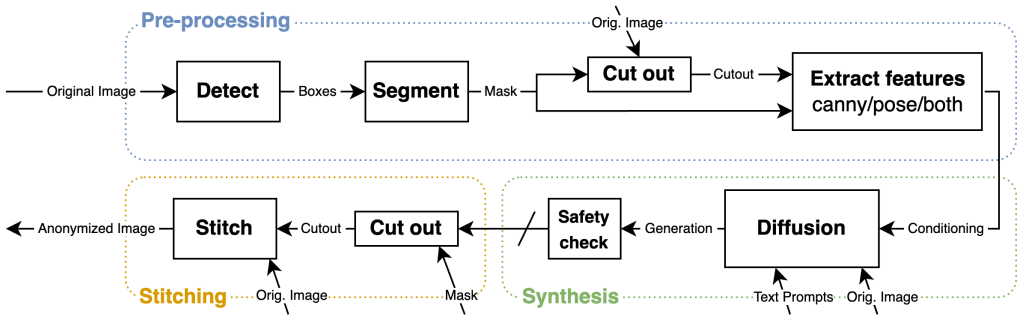


Figure 2: RAD's three-step pipeline and the modules associated with each step.

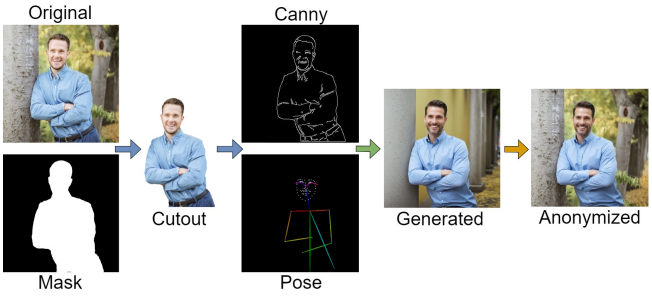


Figure 3: Conceptual overview of our approach.

combined using "zero convolutions", ensuring minimal interference with the underlying diffusion model trained on extensive datasets. This architecture allows ControlNet models to refine diffusion results effectively, even with smaller training datasets, yielding high-quality image generations with reduced noise.

3 Threat Model

Here, we briefly describe the targeted threat model assumed when designing our anonymization framework.

- **Adversaries:** Potential adversaries include (1) malicious entities seeking to de-anonymize the subjects in the images for unauthorized identification purposes, and (2) organizations aiming to exploit personal data without consent, possibly for targeted advertising or adversarial surveillance (e.g., intrusive monitoring without consent).
- **Capabilities:** We assume adversaries have access to: (1) advanced face recognition systems (FRS) for comparing original and anonymized images, and (2) computational resources to process large datasets and perform recognition tasks.

Under this threat model, we want to achieve a good privacy-utility tradeoff, in that the anonymized images should (on average) provide good privacy protection of the individuals appearing in the images, while preserving most of the utility of the original images.

4 RAD's Anonymization Pipeline

Figure 2 illustrates the high-level architecture of RAD's comprehensive three-step anonymization pipeline, detailing each individual module within these steps, and Figure 3 offers a conceptual overview of our approach. The process begins with the *pre-processing* step, where people are detected, and structural features of these people are extracted. Second, in the *synthesis* step, a synthetic image is generated using Stable Diffusion, conditioned by the input features. Finally, in the *stitching* step, the synthesized image is seamlessly integrated with the original, preserving the background while anonymizing the people. We next describe each step in more detail.

4.1 Pre-processing

The pre-processing step detects subjects for anonymization and extracts the features used as conditioning input in the synthesis step (to be performed next). This involves four sequential modules.

Detection Module: The pre-processing step starts with a *detection module*. This module detects anonymization targets by using the latest pre-trained You Only Look Once (YOLO) [30] detection model (YOLOv8), with *yolov8x* used by default. The default model performs well without any additional training but switching it for a higher-performing one is easily done through either the settings (if it is a YOLO-compatible model) or through code modifications which is simplified by the pipeline's modularity. The same is also true for other modules in the pipeline. The detection module outputs detection boxes to the segmentation module, marking the areas of the image that contain people (i.e., anonymization targets).

Segmentation Module: The *segmentation module* uses the detection boxes as input to Segment Anything Model (SAM) [19], which helps create a mask for every anonymization target in the image. These masks are then combined into a single binary mask (flattened) that we use as input to both cutout modules and the feature extraction module. While combining detection and segmentation into a single instance segmentation module is an alternative,

(a) Original (b) Canny edge (c) Silhouette* (d) Pose

Figure 4: Examples of different conditioning input types when using ControlNet. (*) Also using canny edge model.

separating them allows us to leverage SAM's powerful zero-shot segmentation capabilities.

Cutout Module: Next, the *cutout module* uses the segmentation mask to create a single cutout of all the people in the original image. This module is bypassed if the "canny silhouette" setting is used since the silhouette edge can be extracted directly from the segmentation mask.

Feature Extraction Module: The *feature extraction module* is used to extract one to three different types of features from the anonymization targets: (1) an edge map, (2) a silhouette edge map, and/or (3) a pose. These features, illustrated in Figure 4, are later used as conditioning of ControlNet in the synthesis step. The edge map is extracted using Canny edge detection on a cutout of the anonymization targets (excluding the background). The level of detail in the edge map can be controlled by setting a lower and upper threshold for edge detection. The silhouette edge map is also extracted using Canny edge detection but on the binary segmentation mask. Finally, the pose of anonymization targets is extracted using OpenPose[1] (via a Python pip package for auxiliary ControlNet tools). To separate the edge map of the person and the edge map of only the silhouette, we simply refer to them as "Canny" and "Canny silhouette", respectively.

To speed up potential anonymization re-runs with alternative settings, we cache segmentation masks and pose conditioning images with the original input images' hashes as identifiers. This ensures these conditioning images are generated only once per input image, regardless of the number of anonymized copies produced.

4.2 Synthesis

The synthesis step is at the heart of RAD. Here, a new image (of people) is first generated in the *diffusion module* before a safety check is applied using the *safety check module*.

- **Diffusion Module:** The diffusion module makes use of a Stable Diffusion XL image-to-image pipeline with ControlNet, via the Hugging Face Diffusers library [12]. This module takes a conditioning image from the feature extraction module (described above) as input to ControlNet, as well as the original image, and text prompts as input to Stable Diffusion.
- **Safety Check Module:** The generated image then undergoes a *safety check*, which halts the process if the image is inappropriate; otherwise, it proceeds to the stitching step.

[1]https://github.com/CMU-Perceptual-Computing-Lab/openpose

(a) Original (b) 0.3 (c) 0.5 (d) 0.8 (e) 0.99

Figure 5: Example results of using different *strength* values.

(a) Original (b) "Canny" (c) "Pose" (d) "Both"

Figure 6: Example results of varying the ControlNet setting which determines ControlNet's input(s) for the diffusion step. See Figure 4 for input image examples.

Strength and Quality of Anonymization: The privacy level achieved by the anonymization is most affected by the strength parameter while the image quality is most affected by parameter settings such as the number of inference steps, the prompt, and the negative prompt. With the strength parameter controlling how much of the original image is used for the diffusion step, it directly affects the privacy level achieved. Typically, a higher strength results in a more anonymized image, at the risk of generating artifacts. Figure 5 provides a basic example illustrating the impact that the strength parameter has on the anonymization of a person. The number of inference steps impacts quality at the cost of the time taken for image generation. There are also settings for how much ControlNet should affect the final image; e.g., how closely the generated image should follow the canny image. See Appendix A for a more detailed explanation and discussion of these settings.

Impact of ControlNet Conditioning: Using ControlNet, we implement anonymization using three conditioning types: "canny," "pose," or "both", where "both" uses a combination of the first two. In general, we have found that if there are few anonymization targets close to the camera (large in the frame), it is recommended to use the "both" option in combination with using the separate setting "canny silhouette". The pose conditioning alone can lead to unrealistic or distorted results if there are many people close together or if people are further from the camera. Figure 6 shows examples of how "canny" keeps some identifying features from the original image, while "pose" does not correctly match the silhouette of the anonymization target. The option "both" in the figure keeps the correct silhouette by using canny with only the silhouette of the anonymization target, while maintaining the pose by also utilizing the pose conditioning. Considering the level of constraints that each method provides (see Figure 4), it is easy to see why canny can retain more information about the anonymization target, while canny silhouette and pose preserve less identifying information.

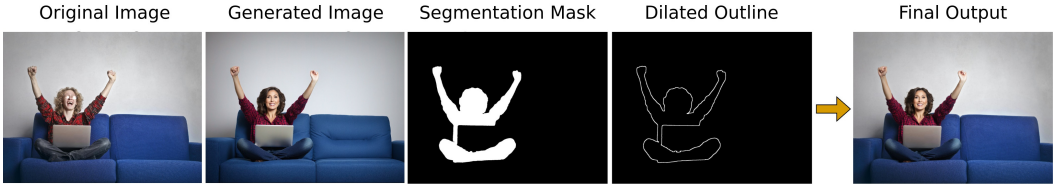


Figure 7: The generated image is stitched onto the original using a segmentation mask and Gaussian blur on the dilated outline.

Figure 8: Naive stitching (left) compared to stitching with blending blurred outline (right).

4.3 Stitching

The stitching step preserves the original background by cutting out the synthesized people and seamlessly stitching them onto the original image. Figure 7 provides an overview of the stitching process, with the detailed steps of each module described next.

Cutout Module: First, the *cutout module* creates a cutout of the synthesized people in the generated image using the same segmentation mask as the earlier cutout module (see pre-processing step). Using the same mask is made possible since people are generated in the same locations as there were people in the original image. The cutout is then passed to the stitching module.

Stitching Module: Next, the *stitching module* blends the cutout onto the original image. A naive approach of simply replacing pixels would leave noticeable jagged artifacts in the image. This is illustrated in the left image of Figure 8. To address this and make the generated image fit in better with the background, we therefore apply a blending process. The blending is done by applying a Gaussian blur (kernel size 5×5) to the outline of the anonymization targets. Here, the pixels that make up the outline of the anonymization targets are determined by reusing the canny silhouette created by the previously described feature extraction module (see pre-processing step). However, to slightly increase the area of effect, the outline is further dilated with a 3×3 kernel size.

While this stitching method does not fully eliminate edge artifacts (e.g., strands of hair sticking out due to inaccurate segmentation, which is generally challenging to fix), we have found the blurring to provide some improvements. Figure 8 illustrates the stitching with and without applying blur.

4.4 Anonymization Settings

RAD supports many different anonymization settings, including controls for the strengths of the anonymization. These settings control the generation via a YAML file that is passed as the first argument to the anonymizer program (as a file path). Appendix A provides details about the different settings and Table 5 summarizes the settings used when generating the results presented here.

Table 1: Primary default settings for each dataset.

Dataset	Strength	Text Prompt	ControlNet Mode
Pexel-Humans	0.75	*people, high quality, neutral lighting, hd, uhd, 4k, 8k*	Both
Cityscapes	0.75	*people, high quality, neutral lighting, hd, uhd, 4k, 8k, light scene, summer day*	Canny

5 Performance Evaluation

5.1 Datasets

Two different datasets are used for our evaluation of RAD. To ensure that we can share results including high-quality images (a scenario typically not considered by prior full-body anonymization papers), a novel dataset, Pexel-Humans, was compiled. This is the primary dataset used in our privacy evaluation. Second, the existing and popular dataset Cityscapes is used for utility evaluation.

Pexel-Humans Dataset: We compiled this dataset from the stock photo website Pexels, with all collected images being licensed under their permissive Creative Commons license. Images were downloaded using the Pexels API with search terms such as "people talking" and "people laughing". Images that did not contain people, had too tiny faces for face recognition, wore face masks, or otherwise were deemed unsuitable for anonymization were removed from the dataset. The resulting dataset consists of 500 images of people with varying gender, ethnicity, and age. Resolution of images varies, but are all bigger than 1024×1024 pixels. Examples images from the dataset are provided in Appendix C.

Cityscapes Dataset: The Cityscapes dataset, designed for evaluating computer vision algorithms in urban scene understanding [8], includes 5,000 high-quality pixel-level annotations for 30 classes and 20,000 weakly annotated images. Only high-quality annotations were used for utility evaluation. Since YOLO models require a specific format, we converted annotations accordingly, generalizing labels like "rider" to "person" and "truck" to "car." Annotations for uncountable items (e.g., "road" or "vegetation") were excluded, focusing on instance-level annotations. Group classifications such as "persongroup," "cargroup", and "bicyclegroup" were also filtered out due to the difficulty of converting them into single instances.

5.2 Example Results: Good vs. Bad

Evaluation Settings: The primary default settings used for the anonymizations of both datasets are provided in Table 1, and the detailed settings are found in Table 5 (in Appendix A). The strength setting 0.75 was used for both datasets since it provided a balance between making enough changes to render subjects anonymous while minimizing any odd artifacts. (We briefly discuss the impact of our parameter selection in Section 6 and provide quantitative example

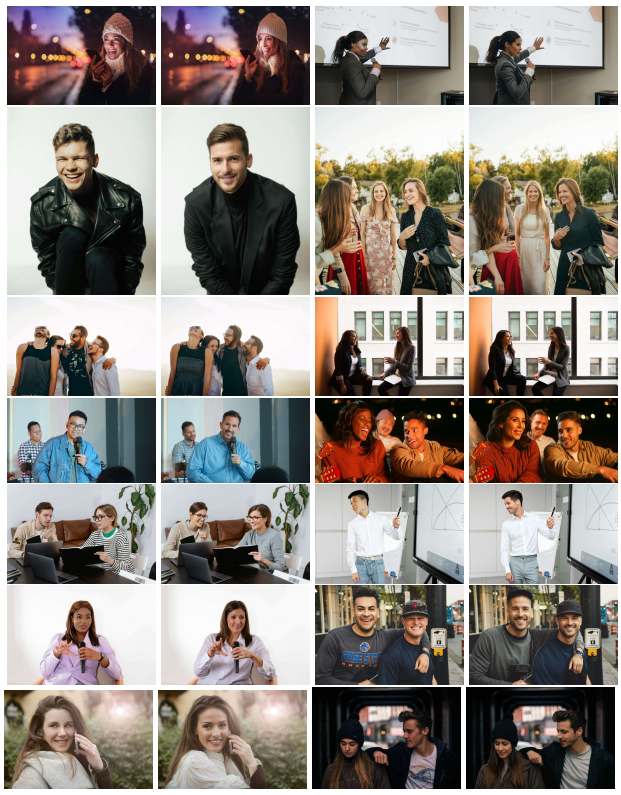


Figure 9: Anonymizations of images from Pexel-Humans. Best viewed by zooming in.

Figure 10: Anonymizations of images from Cityscapes. To the left are original images (with blurred faces to follow Cityscapes licensing) and to the right are realistically anonymized images. Best viewed by zooming in.

comparisons in Appendix B.) The text prompt used was nearly identical, except for the addition of the terms *light scene* and *summer day* for Cityscapes to counter its gray, shadowy lighting conditions, which sometimes resulted in dark, poorly lit synthesized people. As recommended when anonymizing a high-resolution dataset with subjects close to the camera, *both* and *canny silhouette* settings were used for Pexel-Humans. This limited the risk of preserving identifiable details of people close to the camera. For Cityscapes, the far-away anonymization targets allowed the use of *canny* due to the lower risk of extracting identifiable data.

Example Results: Figures 9 and 10 show diverse examples of Pexel-Humans and Cityscapes images before and after anonymization with RAD. The results illustrate that RAD can successfully perform full-body anonymization of people in a wide variety of

(a) Too many fingers or hands. **(b) Poor lighting conditions.**

(c) Incorrect pose.* **(d) Segmentation failure.**

(e) Twins or clones. **(f) Deformed face.**

Figure 11: Observed limitations when anonymizing Pexel-Humans. Original images (cropped) on the left in each subfigure. Best viewed by zooming in. (*) Also segmentation failure.

poses, background contexts, and lighting conditions. It can alter features such as hairstyle, facial features, and clothing. It is also able to handle overlapping people in an image. Although not always the case, gender and ethnicity are also often preserved.

Pexel-Humans is especially interesting since it contains many people who are close to the camera and easily recognizable. Looking at Figure 9, we note that the anonymized subjects are vastly different from their original versions. Yet, everything does not change, as we use the original image as a reference in the diffusion process. However, while clothing often has the same or a similar color after anonymization with the used settings, details that otherwise could be used to identify someone (e.g., tattoos or very specific clothing) are typically relatively generic after the anonymization.

In contrast to Pexel-Humans, images from Cityscapes often contain targets quite far away from the camera. This means that anonymization targets are altered even more with the same strength setting when compared to Pexel-Humans. This is easiest seen by comparing how much more clothes change.

Limitation Examples and Discussion: The generated images are not always perfect and some extra filtering may be needed depending on what the generated images would be used for. To illustrate this, Figure 11 presents some observed limitations when anonymizing Pexel-Humans and Figure 12 shows observed limitations when anonymizing Cityscapes.

For Pexel-Humans, examples were chosen to highlight various challenges in anonymization, particularly those related to the synthesis step, where diffusion tools sometimes generate odd features like extra fingers/hands, turned-away faces, or poor lighting conditions. Naturally, these side effects typically reduce the utility of the image but do not negatively impact the privacy protection. Issues of this kind are time-consuming but, in general, fixable by either generating the image again with a different seed or by changing the text

(a) Missing limbs. (b) Deformed faces.

(c) Hairy faces, and trash can be detected as a person. (d) Odd artifacts and missed person (far right).

(e) Inhuman textile targets. (f) Headless targets.

Figure 12: Observed limitations when anonymizing Cityscapes. Original images (cropped) are to the left in each subfigure. Faces are blurred to follow Cityscapes licensing. Best viewed by zooming in.

prompts to target the unwanted feature specifically. Perhaps more critical challenges are connected to the pre-processing step (typically due to segmentation or detection failures) since they can lead to a target being partly or completely missed during anonymization. Figure 11 (d) shows an example of when segmentation has failed and Figure 14 shows several examples of when detection has failed (counting towards the *missed faces* measure in the face recognition evaluation, presented next). Section 6 discusses how damaged images can be identified and regenerated.

For the Cityscapes dataset, the anonymization targets are often smaller, resulting in a few new challenges, such as creating detailed faces for people far away, or mixing people up with trash cans or bicycles. People are also turned away more than in original images, likely due to difficulties estimating the pose. The impact of these challenges is illustrated in Figure 12.

Finally, we note that RAD only applies anonymizations to people in an image while preserving the background and outline of anonymization targets. This makes RAD practical for anonymizing already annotated images since annotations can be reused after anonymization. However, it also preserves anonymization targets' body shapes, which might make them identifiable in some contexts.

5.3 Privacy Evaluation

We use two methods to evaluate the achieved privacy level: (1) face recognition accuracy when using state-of-the-art face recognition and (2) human evaluation.

Face Recognition to Measure Privacy Level: First, to evaluate face recognition accuracy post-anonymization, we used the state-of-the-art FaceNet512 model via the Deepface library [33] to compare face embedding distances between synthesized faces and their originals, utilizing default cosine distance with a 0.3 threshold.

Table 2: Face recognition results on the anonymized Pexel-Humans dataset. *Cases where the number of faces in the original and anonymized image does not match.

Measure	Explanation	Value
TP ↓	Failed anonymizations	49 (4.5%)
FN ↑	Successful anonymizations	1029 (95.5%)
FP	False recognitions	12
TN	Correctly unrecognized faces	2369
X	Missed faces	25
M	Mismatches*	46

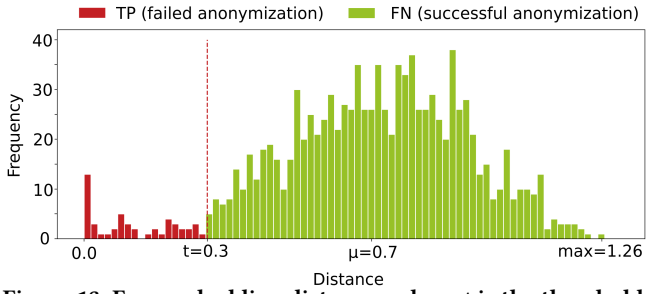


Figure 13: Face embedding distances where t is the threshold for what is considered a recognition and μ is the mean.

This method quantifies RAD's fool rate on a large image set and identifies instances where RAD struggles.

To compare face pairs in Pexel-Humans images, which often contain multiple people, we classified pairs as belonging to the same or different individuals. Pairs were considered the same person if their faces were within 30 pixels of each other in both the original and synthesized images. This method generally works because the pipeline synthesizes people in the same pose as the original images. However, faces that shift more than 30 pixels might lead to misclassifications, affecting recognition accuracy. We carefully monitored recognition results to minimize such errors.

Table 2 presents face recognition results for the anonymized Pexel-Humans dataset and Figure 13 shows the distribution of face embedding distances across all anonymizations (TP + FN). We achieved a 95.5% success rate in fooling FaceNet, based on the ratio of false negatives (successful anonymizations) to total face pairs compared. Among 49 successful recognitions, 25 pairs had distances under 0.1, indicating that the face was completely missed during detection (see examples in Figure 14), and were therefore not properly anonymized. The remaining face pairs are dominated by faces captured from the side, which the recognition model struggles with regardless of how different the faces are (see examples in Figure 15(a)-(b)). Additionally, 46 images (9.2% of images) had some mismatch between the number of detected faces in the original and the synthesized image (see Figure 15(c)). This issue, caused by incorrect face counts, can often be fixed by adjusting generation parameters. A few false negatives, due to segmentation failures, slightly inflated the fool rate.

Human Evaluation on Similarity: We conducted a two-part survey to assess the quality of anonymizations and the ability of humans to identify anonymized subjects. The survey was implemented using Google Forms and representative Pexels images. In total, we had 67 participants, mostly males (79.1%), covering a big

(a) (b) (c)

Figure 14: Examples where a face was missed during the detection step of the anonymization, and therefore did not get properly anonymized. Anonymized Pexel-Humans images (bottom) and the original (upper), with pairwise distances of the faces shown in the bottom row (d:). Here, all faces detected in this step now have a pairwise distance well above the recognition threshold of 0.3, while the missed images all have a pairwise distance close to 0 (suggesting insufficient anonymization for these cases). Here, and in the next figure we use red boxes for pairs that are recognized when using our recognition threshold of 0.3, and green boxes highlight those that are sufficiently anonymized.

(a) (b) (c)

Figure 15: Examples where the recognized faces in profile for which the distance measure is a poor measure of similarity ((a) and (b)) and where there is a mismatch between the number of faces in the original and anonymized images ((c)).

age spectrum (18 to 75+ years). Most participants (68.7%) used a computer, while the rest used a phone.

In the first part, selected the most similar person from five options and rated their similarity, with only one option showing an anonymized version of the target individual.

Each question presented an image of a target person (indicated by a red arrow) and five anonymized images. Respondents selected the image they thought most resembled the target person and rated the similarity on a scale of 1-5 (1 being "very dissimilar" and 5 "very similar"). See Appendix D for example questions.

To generate answer options, we selected an image of the target person in different clothes and pose, then four similar images. All five images were anonymized using RAD. Cutouts of

Figure 16: Original images (top) and anonymized versions (bottom) for question 2 of the human evaluation form. The left-most image is the target person.

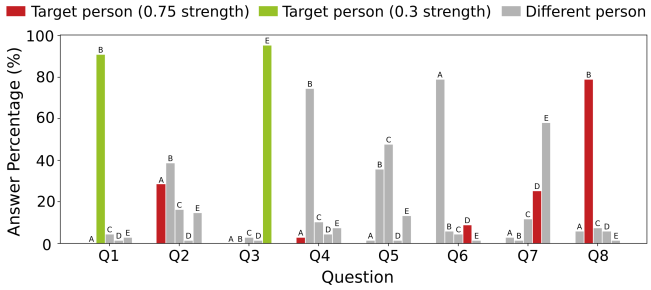


Figure 17: Similarity answers. Colored bars represent the option in which the target person was realistically anonymized. In Q1 and Q3, a lower diffusion strength was used during anonymization. In Q5, the target person was deliberately missing among the options.

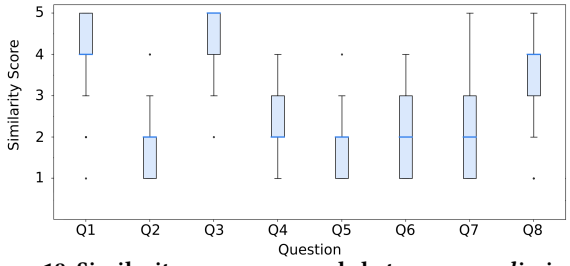


Figure 18: Similarity scores on a scale between *very dissimilar* (1) and *very similar* (5). Blue lines indicate the medians.

the anonymized individuals were placed randomly as options A-E to avoid background influence, and options were shuffled.

Figure 17 shows answer percentages for each of the nine multiple-choice questions, with colored bars indicating the anonymized target options. The tallest bar represents the most popular choice. Figure 18 separately displays each question's similarity score. The bar heights, combined with similarity scores, indicate whether a person was recognized after anonymization.

When interpreting these results, it is important that we included three control questions (Q1, Q3, and Q5). In Q1 and Q3, a lower anonymization strength was used, retaining clothes and accessories, to test if respondents could identify the target person with weaker anonymization. In Q5, the target person was absent to gauge perceived similarity among different individuals. In all other cases, we used an image of the target person in a different pose to ensure the evaluation mimics real-world scenarios where the original image is not publicly available for comparison.

True similarity in Part 1 was measured by the accuracy and confidence of respondents' guesses. High accuracy and confidence indicated *high similarity*, while low accuracy indicated *low similarity*. High accuracy with low confidence also suggested *low similarity*, but less conclusively. Low accuracy with high confidence indicated that another option was more similar than the anonymized target.

Our control questions yielded expected results, indicating survey engagement. First, referring to Q1 and Q3, with lower anonymization strength, we observed over 90% correct picks and high similarity scores (medians of 4.0 and 5.0), regardless of the device used. Second, for Q5, with no target, two options were chosen 36% and 48% of the time, both with a low median similarity score of 2.0. This serves as a baseline for successful anonymization.

For the other (non-control) questions, the results were mixed. The remaining non-control questions had varying results:

- In Q4 and Q6, over 75% of respondents consistently chose an option that was not the target person, with both questions showing a low median similarity score of 2.0. This suggests that respondents did not recognize the target person among the answers. Interestingly, the consistently chosen alternative may indicate which features respondents prioritize. In Q4, the most selected answer matched the target person's age, while the anonymized person appeared younger. In Q6, respondents favored options of the same gender as the target, whereas the anonymized option was of a different gender. This suggests age and gender consistency are significant factors for similarity judgments.
- In Q2 and Q7, respondents rarely chose the target person, with a median similarity score of 2.0. No option in Q2 was picked more than 40%. The even distribution of answers, likely due to similar age and gender among choices, suggests that respondents did not recognize the target person.
- Q8 is notable, with 79% of respondents correctly identifying the target person, with a median similarity score of 4.0. This suggests the anonymization failed in this case, likely due to too many identifiable features being preserved.

Overall, the human evaluation indicates that anonymized images are rarely recognized as similar to the originals. While over 90% of respondents identified the correct individual with low anonymization strength, this majority recognition occurred only once with higher strength. These findings suggest that anonymization strengths do affect recognition, highlighting that people generally cannot identify an anonymized person even among a small group.

Human Evaluation of Achieved Privacy Level: In the second part of the survey, we directly asked participants to rate the effectiveness of anonymizations of eleven image pairs clearly labeled as "Original" and "Anonymized", respectively. Specifically, for each question, respondents were asked to first rate the anonymization effectiveness on a scale of 1-5, with 1 representing "very ineffective" and 5 representing "very effective". Finally, they were asked to rate the photo-realism (evaluated and discussed in the next section) on a scale of 1-5 with 1 representing "very unrealistic" and 5 representing "very realistic". See Appendix D for an example question.

To better understand the impact of image selection and whether facial recognition tools can be used as a good pre-screener, images were chosen based on the face embedding distances gathered during face recognition. Most images, seven out of eleven, were chosen from face pairs with low distances to avoid getting overly optimistic estimates of anonymization effectiveness. The remaining images, four out of eleven, were chosen from face pairs with high distances.

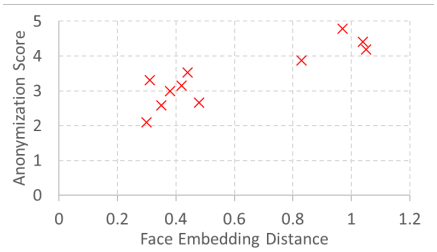


Figure 19: Scatter plot of the mean anonymization effectiveness scores on a scale from *very ineffective* (1) to *very effective* (5) for image pairs with different face embedding distances.

Figure 19 shows a scatter plot of the mean anonymization effectiveness scores on a scale from *very ineffective* (1) to *very effective* (5) for image pairs with different face embedding distances. As a reference point, the mean face embedding distance for all pairs in the dataset is 0.70, predicting an anonymization score of 3.66 based on regression. In general, we observe a positive correlation between face embedding distance and perceived anonymization effectiveness, suggesting that face embedding vectors can be used to estimate the privacy level of different images and sets.

5.4 Utility Evaluation

The utility of the anonymized images was mainly evaluated by (1) using them as training data for an instance segmentation model and (2) asking human evaluators to rate their level of photo-realism. Instance segmentation is a complex task that involves precise detection and pixel-level classification of individual objects. For realistic anonymization, it is important that a person is still recognized as a person, even as their identity is obscured. This task evaluates how well RAD generates new individuals in place of the originals—important for applications like safe navigation, pedestrian detection, patient monitoring, and privacy-conscious customer analysis. We also include some basic object detection results.

Training an Instance Segmentation Model using the Data: Two YOLOv8 segmentation models[2] were trained on the Cityscapes dataset [8]. First, a baseline model was trained on images in their original form. Then, a second model was trained on anonymized versions of images. The two models were then compared in terms of Average Precision (AP) and mean AP (mAP).

Specifically, the models were trained for 100 epochs on the training split of Cityscapes, with early stopping if no improvement was seen within 25 epochs. The evaluation was subsequently done on the validation split. Table 3 summarizes the training configurations: any parameter not present used the default YOLO training settings[3].

Instance Segmentation Evaluation Results: Table 4 presents the segmentation task results, including detection box results (marked as (b) for box and (m) for mask). Comparing the Baseline model's performance on the "original" and the "anonymized" datasets reveals a slight decrease in detection and segmentation accuracy,

[2]https://github.com/ultralytics/ultralytics/tree/v8.2.0
[3]https://docs.ultralytics.com/modes/train

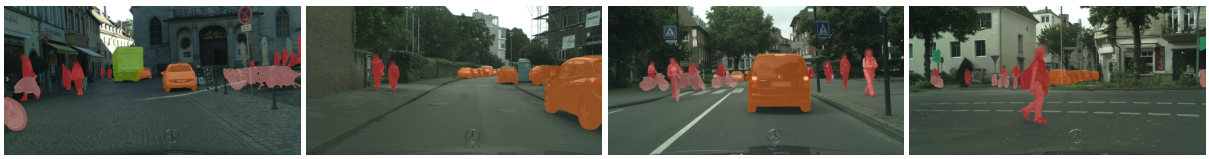


Figure 20: Segmentation examples using a model trained on anonymized Cityscapes. Images are randomly selected images containing humans from the validation set. Faces are blurred to follow Cityscapes licensing.

Table 3: YOLO Training Configuration with Stochastic Gradient Descent (SGD) and early stopping (*).

Parameter	Value	Parameter	Value	Parameter	Value
Model	*yolov8m-seg*	Batch Size	6	Learning Rate	0.01
Epochs	100*	Image Size	1280×640	Weight Decay	0.0005
Patience	25	Optimizer	*SGD*	Overlap Mask	*False*

Table 4: Instance segmentation AP (m) and object detection AP (b) on the Cityscapes dataset with YOLOv8. The results presented are from the best epochs (15 and 32). (*) Early stopping

		Object detection		Instance segmentation	
Dataset	Epochs	$mAP \uparrow (b)$	$AP_{person} \uparrow (b)$	$mAP \uparrow (m)$	$AP_{person} \uparrow (m)$
Original	39* (15)	0.456	0.492	0.422	0.413
Anonymized	49* (32)	0.456	0.477	0.417	0.402

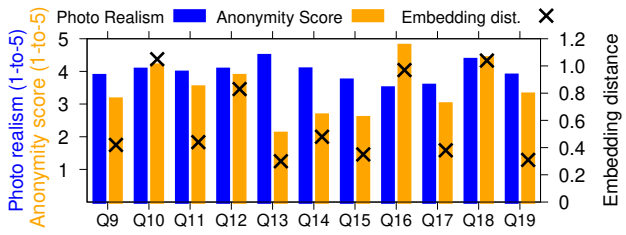


Figure 21: Mean photo-realism scores on a scale from *very unrealistic* (1) to *very realistic* (5). As additional reference points, we also show the mean anonymization scores (scale 1-to-5) and the embedding distances.

particularly for the person class. Despite this, the small decrease indicates that most of the data's utility is preserved. This is further illustrated by example segmentations (e.g., see Figure 20) generated by the model trained on the anonymized Cityscapes.

Human Evaluation on Photo-Realism: As described above, the second part of the human evaluation form (questions 9-19) asks respondents to rate the photo-realism of anonymized images. See Appendix D for an example question.

Figure 21 shows the mean photo-realism score for each question, along with mean anonymization scores and embedding distances. We note that the difference between the lowest and highest mean is only 0.99, meaning that images were considered quite equal in terms of photo-realism. The average score of 3.96 indicates that respondents perceived images to be realistic. This suggests that the photo quality may be sufficient to be utilized as training data for computer vision tasks (e.g., detecting humans in the wild).

No significant trend was observed between photo-realism scores and face embedding distances or anonymization scores (see Figure 21), indicating that better anonymization does not compromise realism. However, notably, despite not seeing the trend for all questions, the images with the lowest and highest photo-realism scores also had the highest and lowest anonymization scores, respectively.

Other Downstream Tasks: We have also run some experiments using YOLOv8, showing that RAD (using default parameters) achieves a strong balance between anonymization and image integrity. For these experiments, we enabled detection of all object classes available in the *yolov8x* model, allowing us to observe how the anonymization process affects various objects within the images. For this evaluation we used three metrics: (1) the average object count per image, (2) the average intersection over union (IoU), calculated as the overlap between detected objects in the original images with their anonymized counterparts, and (3) the average center deviation, calculated as the difference in the central position of detected objects between the original and anonymized images. The first metric allows us to assess the consistency of object detection before and after anonymization, ensuring that the number of detected objects remained stable, the second metric provides an indicator of how well the anonymized images preserved the spatial and semantic integrity of the original content, and finally the third metric gives insight into the positional accuracy retained after anonymization. Collectively they provide a detailed understanding of the model's impact on image quality and the preservation of crucial details necessary for real-world computer vision tasks.

Our results show that the system maintains a high average IoU of 0.88, indicating effective preservation of image semantics. The center deviation is 1.36%, reflecting accurate spatial positioning of anonymized faces. Importantly, the system preserves the number of detected bounding boxes, with a nearly perfect match to the original images (7.0485 vs. 7.0490), ensuring consistent object detection. Additionally, the system does not generate new people, maintaining the integrity of the original image content. Appendix B discusses the impact of the strength parameter on this task.

6 Discussion

Impact of Face Recognition Strength: RAD's privacy-preserving effectiveness depends on the strength of the FRS used. We employed state-of-the-art FaceNet512 models to quantify anonymization by measuring cosine distances between features of original and anonymized faces, supplemented by human evaluations.

While our evaluation is limited to a single FRS, RAD achieved a 95.5% success rate against FaceNet512, demonstrating strong privacy protection. However, as FRS and adversarial techniques evolve, RAD may need ongoing adjustments to maintain its effectiveness.

Privacy-Utility Tradeoff: The adequacy of RAD in protecting privacy against potent adversaries largely depends on maintaining a balance between data utility and privacy. The approach's effectiveness was shown through both automatic face recognition accuracy and human evaluation, confirming its potential in generating high-utility images compliant with privacy regulations. However, ongoing adjustments may be needed as facial recognition technologies and adversarial tactics evolve.

Impact of Strength Parameter: We acknowledge that the selection of the strength parameter (0.75) is preliminary, and the best choice may vary both within and across datasets, depending on factors like the size of the generated person within the image. While higher strength can enhance privacy, it also increases the risk of introducing artifacts due to the limitations of Stable Diffusion. Given the numerous parameters involved (e.g., inference steps, guidance scale, ControlNet scale), it was impractical to exhaustively test all configurations. As such, we acknowledge that tuning may be required for each specific dataset. In Appendix B, we present an evaluation using alternative strength values (0.3, 0.5, 0.99). To better generalize across diverse datasets, interesting future work could be to design methods to dynamically adapt this parameter.

Imperfections of Generated Images: As shown in Sec. 5.2, Stable Diffusion-generated images may contain artifacts and imperfections, a limitation of current generative models. We acknowledge these limitations but believe future advancements will continue to reduce such issues. Our open-source contribution ensures that such enhancements can be easily integrated, improving RAD over time. In the meantime, to improve a dataset, imperfections must be detected and corrected, either through human oversight or automated methods. First, an anomaly detection network can be trained on intentionally damaged images—created via techniques like overlays, noise, or low-step diffusion models—to identify flaws such as segmentation errors and other artifacts. Once detected, regeneration can easily be done by changing the seed or adjusting prompts. In the case that does not work, artifacts can be reduced by lowering the strength setting (at the cost of worse anonymity) or by altering other parameters such as CFG (how closely the generation follows the prompts) and the ControlNet "strength" (how closely the generation adheres to ControlNet input). Figure 22 illustrates how a different seed improves flawed images from Figure 11.

Risk of Generating Real Individuals: The use of pre-trained models like Stable Diffusion presents ethical concerns, including the risk of unintentionally generating images resembling real individuals [7] without their explicit consent. Although our approach aims to create diverse, non-specific human faces, even coincidental resemblances can raise serious ethical issues. Furthermore, explicit consent for inclusion in an original dataset does not imply consent for generating new images. We acknowledge these risks and stress the importance of weighing them against the benefits of producing high-quality anonymized data for different use cases.

High-Quality Datasets: While RAD aims to anonymize individuals to protect privacy, we recognize that high-quality anonymized datasets such as those that RAD can help generate could still be misused in ways that threaten privacy and security, including for privacy-threatening applications like adversarial surveillance or to create convincing fake content for spreading misinformation.

Performance Insights: We also performed a runtime analysis of RAD's pipeline. Key findings are presented here, while a detailed analysis is provided in Appendix F. For example, executing on both a consumer desktop with an Nvidia GTX 2070 GPU and a GPU cluster node equipped with an Nvidia A100 GPU, the average processing time per image was measured across modules. Notably, the desktop, constrained by 8 GB of VRAM, required CPU offloading and experienced slower performance in modules like pose extraction due to single-core operations and lower clock speeds. In contrast, the cluster node's 80 GB VRAM facilitated faster parallelized operations for segmentation and diffusion. However, runtime varied by module, with diffusion affected by inference steps and detection, segmentation, and pose extraction influenced by the number of people in the image. The diffusion module, being particularly resource-intensive, suggests opportunities for optimization to improve efficiency, especially for larger datasets.

7 Related Work

Prior works on realistic anonymization have primarily focused on faces [3, 13, 22] and less on full-body anonymization [14]. However, limiting synthetic data to faces still risks violating privacy since other identifiable features may remain [14].

While GAN-based works like StyleID [22], focusing on facial images, have demonstrated that it is possible to protect individual identities while maintaining the high integrity of dataset characteristics, it is therefore less clear to what extent good utility-privacy tradeoffs can be achieved using diffusion models for full-body anonymization. Only recently have full-body anonymization techniques started being explored to address this gap [5, 14, 16].

RAD's innovations are based on a practical approach to anonymization that covers full-body images and their interaction with the surrounding environment, rather than focusing solely on faces as in the StyleID approach. While StyleID offers a refined technique for feature preservation with a high level of granularity on facial attributes, our method may be more relevant for pedestrian detection, anomymized behavior studies, and other applications benefiting from full-body anonymization.

GANs for Full Body Anonymization: GANs [10] are widely used for various generation tasks, including generating entire images, image in-painting, and super-resolution [2]. Hukkelås et al. presented DeepPrivacy2, a framework for full-body realistic anonymization based on Surface-guided GANs [14, 17]. This method detects, synthesizes, and stitches synthetic human figures and faces, showing promising results for key computer vision tasks [15]. Maximov et al. introduced CIAGANs, generating images using landmarks based on shapes to match poses, although it was mostly evaluated on low-resolution face images [24].

Stable Diffusion for Face Anonymization: Stable Diffusion represents the state-of-the-art in image synthesis, although its use in anonymization has primarily focused on faces. Klemp et al. introduced LDFA, a pipeline using Stable Diffusion for face anonymization [20]. In addition to only anonymizing faces, their work differs in that they process each face individually (by extracting patches) while we synthesize all individuals simultaneously. Since they only use Cityscapes for evaluation (which predominantly contains smaller faces) it is difficult to compare the quality of their anonymizations with ours. Rowan et al. developed a method for face reconstruction using Stable Diffusion conditioned on depth maps with ControlNet, creating a diverse dataset of photo-realistic 3D faces [32]. While their results seem promising and their 3D-mesh approach appears extendable for full-body anonymization, their current work only considers face datasets.

Stable Diffusion for Full Body Anonymization: Most closely related to our work, is the work by Kurzhals [21], who showed that Stable Diffusion can generate slightly modified, synthesized images

(a) Improved hand. (b) Improved lighting. (c) Improved pose. (d) No twins/clones. (e) Fixed face deformation.

Figure 22: Regenerated versions of the imperfect images used to illustrate observed limitations when anonymizing Pexel-Humans (in Fig. 11). Bad example to the left and improved on the right. While we can achieve improved versions of the segmentation failure example too (omitted), the segmentation issues sometimes persist.

to create anonymized versions. Here, we significantly improve on this by prompting Stable Diffusion with cutouts instead of entire images, which are then seamlessly re-incorporated. This method enables more dramatic appearance changes while preserving the original image's integrity, enhancing both privacy protection and data utility.

Video Synthesis: Other researchers have focused on creating temporally consistent videos. For example, OpenAI's SORA [6] uses diffusion model denoising and visual patches to produce varying resolutions and aspect ratios, controlled by textual prompts. Yang et al. [35] introduce a framework that can generate synthetic videos from existing ones that are temporally consistent. Like us they built the tool around Stable Diffusion and ControlNet. However, taking a video-based approach their hardware requirements are substantially higher than ours and background of the video is altered, making it suitable for other use cases than those considered here. Furthermore, on the technical side, they are not concerned with stitching synthesized content back onto original images (possibly reducing the utility of the data) and the framework is not targeting anonymization (which can be seen also by some of their example figures which may be deemed to preserve too much information for effective privacy protection). Building upon this tool, Xia et al. [34] have demonstrated the feasibility of creating an anonymization tool (called DiffSLVA) with the help of this framework. In comparison to these video tools, we have much lower hardware requirements, making our solutions more practical, and we produce higher quality images than what they can achieve for individual frames.

Privacy Evaluation: Prior evaluations of full-body anonymization have shown mixed results regarding privacy. Hanisch et al. [11] highlight limitations in current methods, such as assuming a weak adversary and suggesting more challenging datasets for better approximation of worst-case scenarios. Studies like those by Hukkelas et al. [14] demonstrated successful anonymization with lower re-identification mAP but noted issues with false positives. Klemp et al. [20] found smaller distances between original and anonymized faces using face embeddings, although predominantly evaluating small faces lowered the distances. User studies, such as those by Khamis et al. [18], suggest deepfake obfuscation is promising for privacy and produces more aesthetically pleasing images compared to traditional techniques. Traditional methods like blurring and masking have also been evaluated, with Birnstill et al. [4] and Li et al. [23] noting that masking was more effective for de-identification but less satisfactory in overall image quality. Here, we use a combination of face recognition and human evaluation to demonstrate the level of privacy provided by RAD.

Utility Evaluation: Many studies compare the performance of their realistic anonymization tools to naive techniques due to the lack of standardized baselines [14]. Hukkelås et al. [15] found that their GAN-based tool, DeepPrivacy2, improved instance segmentation and pose estimation tasks compared to blurring, masking, and the earlier version, DeepPrivacy [13]. Similarly, Klemp et al.[20] demonstrated that their Stable Diffusion-based tool, LDFA, outperformed naive techniques and DeepPrivacy models in image segmentation and face detection. Zhou et al. [37] concluded that generative models synthesizing biometric features could mitigate data degradation issues in semantic segmentation tasks. Here, we use a combination of segmentation tasks and human evaluation to demonstrate RAD utility, and note that the generated images in general are of higher resolution than those targeted by most prior works on full-body anonymization.

8 Conclusion

In this paper, we have introduced the Realistic Anonymization using Diffusion (RAD) framework. The RAD pipeline carefully leverages an image-to-image Stable Diffusion model to provide significant advancements in realistic anonymization, effectively reducing the likelihood of recognition while maintaining high image utility. Our evaluations show that RAD provides a high degree of privacy protection, with human assessments aligning closely with face embedding distances in determining the achieved privacy levels. While RAD cannot guarantee complete anonymity due to potential detection and segmentation errors, it significantly mitigates these issues compared to traditional methods. Furthermore, the high utility of RAD is demonstrated by using the anonymized images as training data for instance segmentation tasks and through human evaluations of photo-realism, suggesting their viability for various applications. While some limitations remain (e.g., slower processing times compared to GAN-based models and potential misclassification in diverse lighting conditions), RAD represents a promising approach for anonymizing large datasets while retaining their value for deep learning and other downstream tasks. Future improvements in diffusion models and optimization techniques will likely enhance RAD's effectiveness and efficiency, broadening its applicability and impact. Other interesting future work include evaluating RAD with additional state-of-the-art FRS.

Acknowledgments

We would like to thank our shepherd Vera Rimmer and the anonymous reviewers for constructive comments and feedback that helped improve the paper. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Stability AI. [n. d.]. *Stable Diffusion Launch Announcement.* https://stability.ai/news/stable-diffusion-announcement

[2] Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. 2021. Applications of Generative Adversarial Networks (GANs): An Updated Review. *Archives of Computational Methods in Engineering* 28, 2 (01 3 2021), 525–552. https://doi.org/10.1007/s11831-019-09388-y

[3] Thangapavithraa Balaji, Patrick Blies, Georg Göri, Raphael Mitsch, Marcel Wasserer, and Torsten Schön. 2021. Temporally coherent video anonymization through GAN inpainting. *CoRR* abs/2106.02328 (2021). https://doi.org/10.48550/arXiv.2106.02328

[4] Pascal Birnstill, Daoyuan Ren, and Jürgen Beyerer. 2015. A User Study on Anonymization Techniques for Smart Video Surveillance. In *Proceedings of the 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–6. https://doi.org/10.1109/AVSS.2015.7301805

[5] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. 2017. I Know That Person: Generative Full Body and Face De-identification of People in Images. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1319–1328. https://doi.org/10.1109/CVPRW.2017.173

[6] Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. *Video generation models as world simulators.* https://openai.com/research/video-generation-models-as-world-simulators

[7] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *USENIX Security Symposium*. 5253–5270.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2024. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). https://doi.org/10.1109/TPAMI.2023.3261988

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (10 2020), 139–144. https://doi.org/10.1145/3422622

[11] Simon Hanisch, Julian Todt, Jose Patino, Nicholas Evans, and Thorsten Strufe. 2024. A False Sense of Privacy: Towards a Reliable Evaluation Methodology for the Anonymization of Biometric Data. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2024 (2024), 116–132. Issue 1. https://doi.org/10.56553/popets-2024-0008

[12] Huggingface. [n. d.]. *Diffusers.* https://huggingface.co/docs/diffusers/en/index

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*. Springer, 565–578. https://doi.org/10.1007/978-3-030-33720-9_44

[14] Håkon Hukkelås and Frank Lindseth. 2022. DeepPrivacy2: Towards Realistic Full-Body Anonymization. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. https://doi.org/10.1109/WACV56688.2023.00138

[15] Håkon Hukkelås and Frank Lindseth. 2023. Does Image Anonymization Impact Computer Vision Training?. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 140–150. https://doi.org/10.1109/CVPRW59228.2023.00019

[16] Håkon Hukkelås and Frank Lindseth. 2024. Synthesizing Anyone, Anywhere, in Any Pose. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4023–4034. https://doi.org/10.1109/WACV57701.2024.00399

[17] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. 2023. Realistic Full-Body Anonymization with Surface-Guided GANs. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1430–1440. https://doi.org/10.1109/WACV56688.2023.00148

[18] Mohamed Khamis, Habiba Farzand, Marija Mumm, and Karola Marky. 2022. DeepFakes for Privacy: Investigating the Effectiveness of State-of-the-Art Privacy-Enhancing Face Obfuscation Methods. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI)*. Article 21, 5 pages. https://doi.org/10.1145/3531073.3531125

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

[20] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. 2023. LDFA: Latent Diffusion Face Anonymization for Self-driving Applications. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3199–3205. https://doi.org/10.1109/CVPRW59228.2023.00322

[21] Kuno Kurzhals. 2023. Privacy in Eye Tracking Research with Stable Diffusion. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications (ETRA)*. Article 70, 7 pages. https://doi.org/10.1145/3588015.3589842

[22] Minh Ha Le and Niklas Carlsson. 2023. StyleID: Identity Disentanglement for Anonymizing Faces. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2023 (2023), 264–278. Issue 1. https://doi.org/10.56553/popets-2023-0016

[23] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1343–1351. https://doi.org/10.1109/CVPRW.2017.176

[24] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5446–5455. https://doi.org/10.1109/CVPR42600.2020.00549

[25] Ricardo Mendes and João P. Vilela. 2017. Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access* 5 (2017), 10562–10582. https://doi.org/10.1109/Access.2017.2706947

[26] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. 2016. MOT16: A Benchmark for Multi-Object Tracking. *CoRR* abs/1603.00831 (2016). https://doi.org/10.48550/arXiv.1603.00831

[27] Council of European Union. [n. d.]. *Regulation (EU) 2016/679 of the European Parliament and of the Council.* https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[28] Anastasiia Pika, Moe T. Wynn, Stephanus Budiono, Arthur H.M. ter Hofstede, Wil M.P. van der Aalst, and Hajo A. Reijers. 2020. Privacy-Preserving Process Mining in Healthcare. *International Journal of Environmental Research and Public Health* 17, 5 (2020). https://doi.org/10.3390/ijerph17051612

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2307.01952

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788. https://doi.org/10.1109/CVPR.2016.91

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[32] Will Rowan, Patrik Huber, Nick Pears, and Andrew Keeling. 2023. Fake It Without Making It: Conditioned Face Generation for Accurate 3D Face Reconstruction. *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2307.13639

[33] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *Proceedings of the 2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

[34] Zhaoyang Xia, Carol Neidle, and Dimitris N. Metaxas. 2023. DiffSLVA: Harnessing Diffusion Models for Sign Language Video Anonymization. *arXiv preprint* (2023). https://doi.org/10.48550/arXiv.2311.16060

[35] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. In *Proceedings of SIGGRAPH Asia 2023*. Article 95, 11 pages. https://doi.org/10.1145/3610548.3618160

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355

[37] Jingxing Zhou and Jürgen Beyerer. 2022. Impacts of Data Anonymization on Semantic Segmentation. In *Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV)*. 997–1004. https://doi.org/10.1109/IV51971.2022.9827262

A Anonymization Settings

General Settings: These are basic settings such as input and output folder paths, whether an output folder should be overwritten, and in which format anonymized images should be saved.

- *Input and output* folder paths specify the path to a folder of images that should be anonymized and the path to which the anonymized images should be saved.
- *Overwrite* is a setting that toggles if any existing folder at the output path should be overwritten. If set to false RAD will skip all already generated images and only generate images that have not yet been processed.

- *Save format* specifies in what image format the output images should be saved. The default is to use the same format as the input image.

Diffusion Settings: The settings that have the most significant impact on the quality of the generated persons are strength, inference steps, prompt, and negative prompt. Strength generally controls how anonymous the final image will be, at the risk of generating artifacts. The number of inference steps impacts quality at the cost of the time taken for image generation. There are also settings for how much ControlNet should affect the final image, e.g. how closely the generated image should follow the canny image. Each setting is described in more detail below.

- *Strength* controls how much the original input image should be changed, i.e. how much "freedom" the Stable Diffusion model has in the generation. A higher value means that the generation is allowed to stray further from the original image contents. This means that the strength setting can be seen as a degree of anonymization, with a higher value usually correlating to a more anonymous person. Too high strength values, however, tend to produce unrealistic results with artifacts such as too many arms or fingers. Also, a lower strength can sometimes produce a very anonymous generation for one input image, but a not-so-anonymous generation for another. A further observation is that people who occupy a larger part of the frame usually require a higher strength setting to be as anonymous than a person who occupies a smaller part of the frame. See Figure 5 for an example of how strength affects the anonymization of a person. The strength should be set between 0-1, and it is recommended to use a value of around 0.75 if people are close in the frame.
- *Prompt* and *negative prompt* are conditioning textual inputs to Stable Diffusion used in the diffusion process. The prompt specifies what is wanted in a generated image, and the negative prompt specifies what is unwanted. When, for example, trying to achieve photo-realism with natural lighting conditions, the text prompt could include terms such as *high quality*, *hd*, *4k*, and *neutral lightning*. The negative prompt could include terms like *cgi*, *3d*, *drawing*, and *sunset*. Important to note is that some artifacts in an image are hard to fix only by adjusting the prompts, such as blurry images or clones (the same person generated multiple times). Such artifacts are usually fixed by using higher-resolution images. Also worth noting is that the text prompts are separated from the input image. Therefore, instructional prompts such as "change the shirt to a red jacket" do not have the desired effect.
- *Inference steps* control how many steps the diffusion denoising process should take. A higher number of steps usually results in higher quality image generations but comes at the cost of time. Hugging Face Diffusers will use fewer inference steps than specified at a lower strength according to $\textit{inference steps} = \frac{1}{\textit{strength}}$. A recommended setting for inference steps is between 15 and 50.
- *ControlNet conditioning scale* defines how much ControlNet will affect the generated image. A higher value means that the generated image more closely aligns with the ControlNet conditioning input (canny or pose images). This setting should be set in the range of 0-1, but can be set to higher at the risk of producing artifacts. A recommendation is to set this to close to 1.
- *Guidance scale*, also known as CFG, controls how much of an effect the prompts have on the generated image. A higher value means that the prompts have a higher impact on the image generation. This value should be in the range of around 0-50, but a good recommendation is to keep it around 7.5.
- *Seed* controls what seed to use for the randomness in the noise added during the diffusion process. This setting could be left unset and will then default to a random seed. The main benefit of setting the seed manually is to get predictable results. Anonymizing an input image with identical prompts, strength, conditioning, guidance scale, and seed will result in the same generated image.

ControlNet Settings: These settings specify what ControlNet conditioning should be used and how the conditioning input should be extracted from the original image.

- *ControlNet mode* controls the type of ControlNet conditioning that will be extracted from the original image and used as conditioning in the image generation. This option can be set to either "canny," "pose," or "both". The options "canny" and "pose" will only extract and use one of these conditioning types, while "both" will use a combination of the two. In general, if there are few anonymization targets in an image that is close to the camera (large in the frame), it is recommended to use the "both" option in combination with using the separate setting "canny silhouette". The pose conditioning alone can lead to unrealistic or distorted results if there are many people close together or if people are further from the camera.
- *Canny minimum and maximum thresholds* are specific settings for the canny extraction process. These settings determine how much detail is extracted from the original image of the anonymization target. Higher values for the thresholds will result in a less detailed canny conditioning image, effectively providing higher anonymity. However, values that are too high defeat the purpose of extracting edges. Generally, the minimum and maximum thresholds should keep a ratio of between 1:2 and 1:3 for optimal results. Another important aspect is that the selected thresholds might affect varying images differently. For example, canny detects more edges for sharp images with high contrast. Therefore, the threshold might have to be adjusted depending on the types of images in a dataset. Because of the troublesome process of finding a threshold suitable for an entire dataset, canny edge detection is not ideal for the anonymization task. However, some kind of edge detection is often necessary, especially for smaller anonymization targets in the frame where the pose is harder to determine.
- *Canny silhouette* is a toggle setting that tries to solve most anonymization issues with the regular canny extraction process. If toggled on (set to true) the canny conditioning will only be used with a canny image that contains extracted silhouettes of anonymization targets (by extracting the canny

(a) Original image **(b) Realistic Vision XL** **(c) SDXL Base** **(d) SDXL Base + Refiner**

Figure 23: Result of using different diffusion models.

conditioning image from the segmentation mask as opposed to from the original image cutout). This means that no information about a person's appearance outside of their silhouette (and possibly their pose) is used for the ControlNet conditioning, resulting in a higher guarantee of privacy. The original image is still used as input to the generation process, so a high strength value is still required for high anonymization.

Model Settings: All models used in the pipeline can be changed through the settings. This includes the main diffusion model, ControlNet conditioning models, and the models used for creating segmentation masks. The choice of models that affect anonymization the most is the diffusion-related models and the object detection model. All models can be defined as local paths, which can be important when, for example, running an anonymization in a containerized setting. The main diffusion model and ControlNet models can also be defined as Hugging Face repository names such as "stabilityai/stable-diffusion-xl-base-1.0". Similarly, the YOLO and SAM models will be automatically downloaded if only specified as the name of the model.

- *Diffusion model* is the setting that defines the main Stable Diffusion model used to generate images. Without code modification, the only valid models are Stable Diffusion XL models that support image-to-image generations. Example results with different models are shown in Figure 23.
- *Refiner model* is an optional setting that defines the refiner model to be used. A refiner is a specific diffusion model that increases the quality of an already generated diffusion image. If this setting is left out, no refiner will be used. The necessity of a refiner depends on the quality of the diffusion model output.
- *ControlNet canny model* defines the model used for canny conditioning images. This should be a model that is trained to specifically condition image generation using canny conditioning images.
- *ControlNet pose model* defines the model used for pose conditioning images. This should be a model that is trained to specifically condition image generation using pose conditioning images.
- *YOLO object detection model* specifies the YOLO model that is used to detect anonymization targets in an image. The resulting detection boxes are then used as input to SAM for creating segmentation masks. This chosen detection model needs to be compatible with YOLO, if no modification is made to the code. The choice of model is very important for anonymization purposes since it detects the targets (people) to anonymize. A model with poor detection performance will miss people in an image. This in turn results in people from the original image not being anonymized and instead kept as is, just as the background of the original image is preserved.
- *SAM segmentation model* specifies what SAM model to use for segmentation. This has to be a SAM-compatible model, unless the code is modified.

Optimization Settings: An issue with Stable Diffusion (and especially Stable Diffusion XL) is that it, by default, uses a large amount of VRAM when generating images. This is a problem when generating images on a consumer GPU since these usually do not have more than 8 GB of VRAM. A solution for this is offloading, i.e. moving the models between the GPU and CPU.

- *Maximum output width and height* specify the maximum size of the input image. If an input image exceeds these dimensions, it will be resized to fit. Additionally, all input images will be cropped to be a multiple of 8 since this is a requirement for the Hugging Face Diffusers. Larger input images will result in larger generated images (same dimension as the input), requiring more VRAM and resulting in longer image generation times. Therefore, an important optimization can be to adjust the image dimensions. Note that input images with a width or height below 512 pixels can result in artifacts in the generated image.
- *CPU offloading* specifies the amount of offloading used. This setting accepts values from 0-3, where zero represents no offloading, and three represents maximum offloading. A value of one means that whole models are offloaded, a value of two results in sub-models being offloaded, and a value of three means that sub-model offloading and VAE tiling are employed. More offloading reduces the amount of VRAM used but comes at the cost of additional execution time for image generation.
- *Compile U-Net* specifies if the U-Net used in the Stable Diffusion generation should be compiled or not (a true or false value). A compiled U-Net can significantly speed up image generation time, especially if a large batch of images is anonymized. However, this can only be used without CPU offloading (offloading with a value of zero).

Device Settings: Device settings specify which hardware device to use for image generation and creation of segmentation masks. The segmentation setting also specifies if segmentations should be run in parallel or sequentially.

- *Device* specifies what device the images should be generated on (what device to run Stable Diffusion on). This should, for example, be set to "cuda" if running on an Nvidia graphics card.
- *Segmentation* is a setting that specifies how the segmentation should be run. By default, segmentation masks will be created in parallel on the same device as the image generation. This results in almost no additional time being required to create segmentation masks after the first image. An option if VRAM resources are limited is to set segmentation to run in parallel on, for example, the CPU instead. An additional

Table 5: Complete anonymization settings for Pexel-Humans (PH) and Cityscapes (C).

Diffusion Settings	
Strength	0.75
Prompt	*people, high quality, neutral lighting, hd, uhd, 4k, 8k, light scene (C), summer day (C)*
Negative prompt	*(deformed iris, deformed pupils, semi-realistic, cgi, 3d, render, sketch, cartoon, drawing, anime, extra face, clone, cloned face), no pants, no shirt, swimwear, lightly dressed, sunrise, sunset, lamp, bright light, text, cropped, out of frame, worst quality, low quality, jpeg artifacts, duplicate, morbid, mutilated, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck, UnrealisticDream*
Inference steps	53
ControlNet scale	0.9
Guidance scale	7.5
Seed	0
ControlNet Settings	
Mode	*both (PH) canny (C)*
Canny thresholds	100/200
Canny silhouette	*true (PH) false (C)*
Model Settings	
Diffusion model	$RealVisXL\_V3.0$
Canny model	$controlnet - canny - sdxl - 1.0$
Pose model	$controlnet - openpose - sdxl - 1.0$
YOLO model	$yolov8x$
SAM model	sam_vit_h

option, if experiencing other hardware limitations, is setting segmentation to "seq," meaning that the segmentation mask is created sequentially with each input image.

Settings Used for Experiments: Table 5 presents the complete anonymization settings for both the Pexels-Human and Cityscapes datasets used for privacy and utility evaluations.

Discussion of Settings and their Tradeoffs: RAD comes with many user-controllable parameters. This allows flexibility in the anonymization of varying types of images. It also gives the user a lot of control over what is generated. However, finding settings suitable for anonymizing large datasets can be both difficult and time-consuming. Many parameters directly impact the privacy-utility tradeoff and can tip the scale one way or the other in terms of identifiability. Following are discussions about the settings that are particularly important, since these have the greatest effect on the quality of anonymizations:

- *Strength* is perhaps the most intuitive setting, and also one that has an immediate impact on anonymity. Higher strength leads to better anonymizations, which is reflected in the results. But a higher strength setting can also lead to artifacts in a generated image, and increased generation time if one wants to keep the same level of detail. Finding a one-size-fits-all value for this setting would be preferable, but the required strength is highly dependent on the context. For instance, the strength setting will alter subjects further away from the camera more than close subjects. This can be seen when comparing the anonymized images from Cityscapes with those from Pexel-Humans since the datasets vary in this regard (see Figure 10 and Figure 9). Both these datasets were anonymized using 0.75 strength.
- *ControlNet mode* also has a significant impact on anonymity, since it controls what features are extracted from people in an image. From a utility perspective, the more features that can be preserved, the better; however, from a privacy perspective, the opposite is true. The available choices in RAD each have pros and cons. Canny edge maps are particularly useful for preserving pose and blending anonymization targets well into the background. However, edge maps also run the risk of preserving identifiable information in the image, especially when anonymization targets are large in the frame. Pose maps preserve pose with less information than edge maps, but perform poorly for anonymization targets that are small in the frame. The recommended *both* option achieves the most consistent results by using pose with only the silhouette edges using *canny silhouette*. This option reduces the risk of preserving identifiable features while enabling reliable pose replication.
- *Detection, segmentation, diffusion*, and *ControlNet models* all have a fundamental impact on anonymizations. This is a consequence of constructing a pipeline with several existing models; the anonymization quality is limited by the performance of each one. The detection and segmentation models are critical for finding and producing masks for every anonymization target in an image, greatly affecting the level of achieved privacy protection. The Stable Diffusion and ControlNet models affect both the quality and photo-realism of generated images and the achieved privacy level.
- *Text prompt* and *negative text prompt* can be varied endlessly and are central to Stable Diffusion. They can control both general and detailed aspects of image generation. General text prompts that fit most images are best suited for anonymizing an entire dataset with diverse images. This thesis has focused on achieving prompts that work reasonably well for the datasets used, instead of finding optimal prompts. So-called "prompt engineering" is a novel field of research, born from the desire to design optimal prompts, and deserves more attention in future work.

B Impact of the Strength Parameter

To better understand the impact of the strength parameter, we ran RAD with different strength parameters in the Pexel-Human dataset.

High-level Results: Table 6 summarizes these results for the strengths 0.3, 0.5, 0.75, and 0.99. We note that RAD achieves the best performance at with a strength of 0.75; the default choice used in all reported experiments. With this setting, we achieve a 95.5% successful anonymization rate (1,029 FN out of 1,078 total faces) and only 12 FP. This setting outperforms lower strengths (0.3 and 0.5), which show higher failed anonymizations (200 and 97 TP, respectively), while avoiding the over-anonymization seen at 0.99 strength (36 FP). In general, 0.75 results in the most successful anonymizations (FN) and the most faces correctly classified as different (TN), as well as the fewest failed anonymizations (TP) and

Figure 24: Samples images from the Pexels-Humans dataset.

Figure 25: Sample Images from Cityscapes dataset with segmentation masks for object instances (e.g. red=person, orange=car, pink=bicycle, green=traffic light). Faces are blurred to follow Cityscapes licensing.

Table 6: High-level summary results using different strengths parameter settings with the Pexel-Human dataset. All other parameter values remain the default values.

		Strength parameter			
Measure	Explanation	0.3	0.5	0.75	0.99
TP ↓	Failed anonymizations	200	97	49	56
FN ↑	Successful anonymizations	932	1,023	1,029	962
FP	False recognitions	20	24	12	36
TN	Correctly unrecognized	2,366	2,362	2,369	2,288
X	Missed faces	45	33	25	23
M	Mismatches*	23	32	46	85

Table 7: Object detection results using different strengths parameter settings with the Pexel-Human dataset together with YOLOv8. All other parameter values remain the default values.

	Strength Parameter			
Parameter Value	0.3	0.5	0.75	0.99
Avg Object Count	6.9556	6.9839	7.0485	6.8508
Avg IoU	0.9134	0.8979	0.8751	0.8477
Avg Center Deviation	0.91%	1.12%	1.36%	1.77%

false recognitions (FP). Overall, the intermediate strength of 0.75 appears to effectively balance privacy preservation and data utility.

Object Detection Results: To evaluate the impact of the strength parameter of RAD on image quality and object detection performance, we re-ran our YOLOv8 experiments from Section 5.4 for each of the considered parameter choices. Table 7 summarizes these results.

Our face anonymization system shows varying performance across different strength parameters, reflecting the tradeoffs between anonymization and image quality. As the strength parameter increases, the average Intersection over Union (IoU) decreases, from 0.91 at a lower strength to 0.88 at 0.75, and 0.85 at the highest setting, suggesting that higher strengths may lead to over-anonymization, potentially compromising image semantics. The center deviation, which measures the positional accuracy of detected objects, also increases with higher strength, rising from below 1.12% at lower strengths to 1.36% at 0.75, and 1.77% at the highest considered strength.

The 0.75 strength represents a balanced choice, offering enhanced anonymization while maintaining reasonable spatial accuracy and image quality. At this setting, the system nearly perfectly preserves the number of bounding boxes (7.0485) relative to the original images (7.0490), ensuring consistent object detection. Additionally, the system does not generate new individuals, maintaining the integrity of the original content. By excluding unpaired boxes from the center point deviation and IoU calculations, we ensure an accurate assessment of the system's performance on successfully anonymized faces.

This analysis highlights the importance of carefully selecting the strength parameter to balance privacy with the retention of essential image details, particularly in applications requiring high utility from the anonymized data.

C Example Images from the Datasets

Figure 24 example images from our Pexels-Human dataset.

Figure 25 shows two examples from the Cityscape dataset, including also the corresponding segmentations.

D Example Questions from Survey

Figure 26 shows two example questions used in the first part of the user survey. Again, three of these questions (Q1, Q3, and Q5) are control questions, while the others are regular questions (Q2, Q4, Q6, Q7, Q8). Figure 27 shows an example of the question type used in the second part of the survey (Q9-Q19).

E Licensing Information

The pipeline makes use of several existing tools with a few different licenses. Licenses and basic license information for all major tools are provided in Table 8. Most licenses for the tools used in this pipeline are very permissive, but some prohibit commercial use or mandate that any derived programs must disclose their code as open source.

F Example Runtimes and Bottlenecks

The average execution time for each module was measured twice: once on a consumer desktop and once on a GPU cluster node. The graphics card in the consumer desktop was an Nvidia GTX 2070 with 8 GB of VRAM, while the cluster node had an Nvidia A100 with 80 GB of VRAM. For these measurements, 10 sample images from the Cityscapes dataset were chosen at random. These images have an image size of 2048 × 1024 pixels (PNG format) and contain varying amounts of people (from none up to 20 people). For these settings, both canny and pose were used, and an empty cache folder, meaning no segmentations or pose images were cached

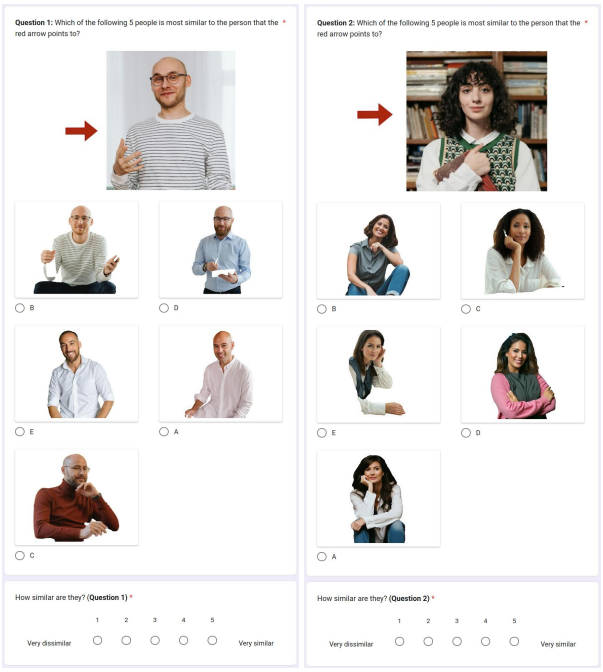


Figure 26: Sample questions from first part of the user survey. A comparison of original images and anonymized versions for question 2 (right image) can be found in Figure 16.

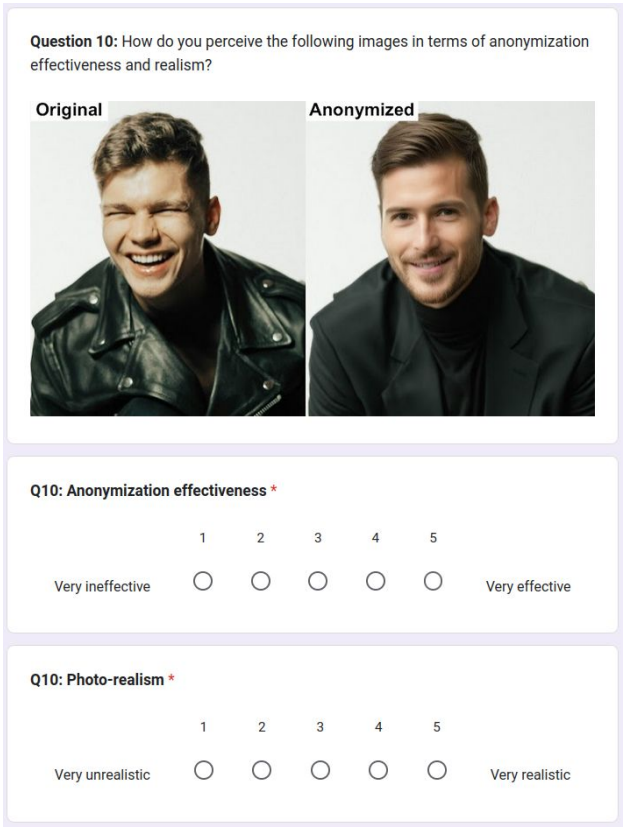


Figure 27: Sample question from second part of the survey.

Table 8: Licensing information for each tool, including license requirements on copyleft and non-commercialization. Many of these licenses also prohibit the use of the author(s) name in promotional material for derived products.

Tool	Type	License	Closed-source	Commercial
YOLO	ML Framework	AGPL-3.0	✗	✓
ControlNet Aux.	Preprocess	Apache-2.0	✓	✓
OpenPose	Preprocess	Custom	✓	✗
HF Diffusers	Diffusion	Apache-2.0	✓	✓
DeepFace	Face recognition	MIT	✓	✓
SAM	Preprocess	Apache-2.0	✓	✓
PIL (Pillow)	Image processing	HPND	✓	✓
NumPy	Image processing	Custom	✓	✓
PyTorch	ML Framework	Custom (BSD)	✓	✓
Matplotlib	Graphs	PSF	✓	✓

Table 9: Runtime measurements (in seconds) for each module in the anonymization pipeline. Notes: (*) Executed in a thread separate from the rest of the modules. (**) Only runs on the CPU in the current implementation.

Module	Cluster Node	Consumer Desktop
Detection	0.46	1.30*
Segmentation	0.46	26.16*
Cut out	< 0.1	< 0.1
Canny extraction	< 0.1	< 0.1
Pose extraction**	59.7	15.3
Diffusion (39 steps)	15.3	182
Safety Check	< 0.1	< 0.1
Stitch	1.1	0.95
Total	**77.3**	**203**
Total w/o pose	**17.6**	**187**

beforehand. The optimization settings were set to maximize VRAM usage without running out of memory.

Table 9 summarizes the average time it took to anonymize one image, measured for every module in the pipeline. The time for the first image was disregarded since it included extra initialization time that becomes irrelevant when anonymizing a more extensive set of images.

When interpreting these results, it should be noted that the limited VRAM (8 GB) of the consumer desktop meant that it required maximum CPU offloading, in turn preventing the U-Net from being compiled. The limited VRAM also meant that the parallel segmentations were run on the CPU. In contrast, the larger VRAM of the cluster node (80GB) ensures that no CPU offloading was needed, we could compile the U-Net and the segmentation could be performed sequentially on the GPU.

In general, the key factors that most influence the runtime are distinct for each module. For example, the runtime for diffusion is primarily influenced by the number of inference steps used, while the runtimes for detection, segmentation, and pose extraction are predominantly influenced by the number of people in an image.

Some operations running on the CPU are slower on the cluster than on the consumer desktop. This is despite having many more CPU cores available on the cluster. However, the consumer desktop cores run at a higher clock frequency than those on the cluster. Poorly parallelized operations can only run on one or a few cores

at a time, resulting in clock speed being the main runtime bottleneck. This can be seen in the pose extraction module, which is not parallelized in the current implementation, resulting in almost a four-time increase in runtime. Conversely, parallelized operations such as segmentation and diffusion run extremely fast on the cluster. This is partly due to the cluster GPU having more cores and faster clock speeds than the desktop GPU.

The usage of maximum CPU offloading on the consumer desktop significantly impacts the diffusion runtime, since this requires the CPU to offload parts of the diffusion pipeline when not in use. This also affects the parallel segmentation execution time and vice-versa since these operations were also run on the CPU.

Overall, the pipeline has a few limitations, primarily related to the runtime of diffusion, object detection, and the pose detector. First, while the diffusion models in RAD provide high-quality and controlled image generation, they are more resource-intensive and slower than GAN-based alternatives, especially with limited VRAM. Optimizing runtime and memory usage could make diffusion-based anonymization more feasible for large datasets. Second, RAD currently relies on object detection for anonymization targets. Adding a face detector in the pre-processing and verification stages could enhance detection accuracy and privacy by ensuring faces are properly anonymized and re-generated if needed. Third, the pose detector, using a community OpenPose implementation, often misses people and runs on the CPU, increasing runtime. Using the official OpenPose Python API could improve performance but might complicate installation. Our current implementation is relatively easy to install and run on different systems.