On the importance of conditioning for privacy-preserving data augmentation

Julian Lorenz, Katja Ludwig, Valentin Haug, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Lorenz, Julian, Katja Ludwig, Valentin Haug, and Rainer Lienhart. 2025. "On the importance of conditioning for privacy-preserving data augmentation." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, October 19–23th, 2025, Honolulu, Hawai'i, 2317–26. Piscataway, NJ: IEEE. https://arxiv.org/abs/2504.05849.

On the Importance of Conditioning for Privacy-Preserving Data Augmentation

Julian Lorenz Katja Ludwig Valentin Haug Rainer Lienhart
University of Augsburg
Augsburg, Germany
{julian.lorenz,katja.ludwig,valentin.haug,rainer.lienhart}@uni-a.de

Abstract

Diffusion models can be used as a powerful augmentation method to artificially extend datasets for enhanced training, by swapping parts of the image with generated content which look very different from the originals to the human eye. A recent ECCV oral paper [15] has suggested to use this data augmentation technique for data anonymization. However, we show in this paper that diffusion models that are conditioned on modalities like depth maps or edges to guide the diffusion process are not suitable as a privacy-preserving method. We use a contrastive learning approach to train a model that can correctly identify people out of a pool of candidates, with a success rate of over 69% regarding a pool of over 3k persons. Moreover, we demonstrate that anonymization using conditioned diffusion models is susceptible to black box attacks. We attribute the success of the described methods to the conditioning of the diffusion model in the anonymization process. The diffusion model is instructed to produce similar edges for the anonymized images. Hence, a model can learn to recognize these patterns for identification.

1. Introduction

In recent years, neural networks have grown significantly in size. However, training large models effectively requires extensive datasets to prevent overfitting. While advancements in pre-training on massive unlabled datasets have reduced the need for massively labeled datasets for fine-tuning, a substantial amount of data is still necessary. This poses a challenge for machine learning tasks involving humans, such as human pose estimation or face recognition, where collecting large-scale datasets is difficult. Obtaining comprehensive consent from all individuals in custom datasets is often impractical, further complicating data acquisition.

A potential solution to this challenge is privacy-preserving data augmentation, which aims to obscure individuals' identities by replacing the parts of images showing real humans with synthetically generated images while keeping the background. This approach preserves the diversity of real-world (in-the-wild) datasets while ensuring anonymity. However, these models are not entirely foolproof and may still leak sensitive information. In the worst

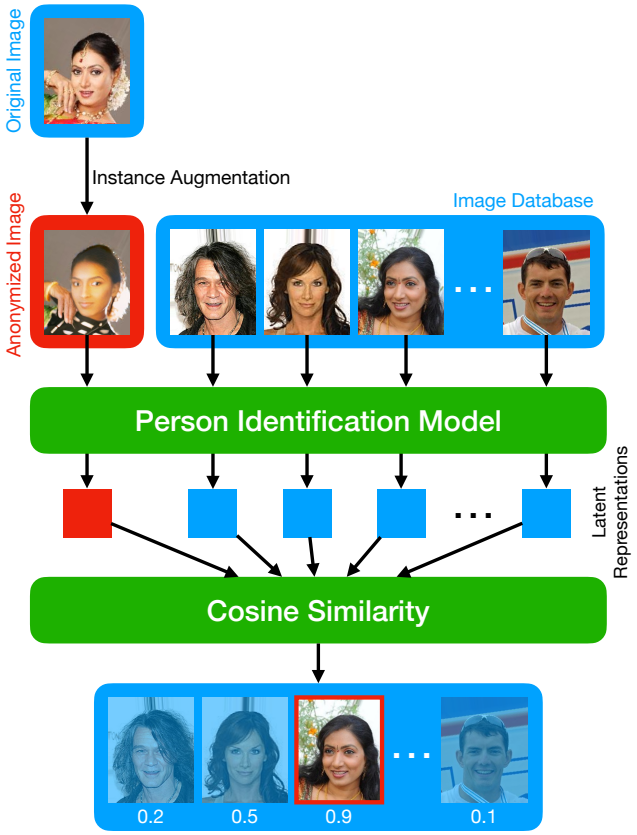

Figure 1. Overview of our method. An original image is anonymized with Instance Augmentation. Our model takes the anonymized image and an image database as input and outputs latent representations for all images, encoding the person identity. We calculate the cosine similarity scores for the anonymized representation vector and the representations of all persons in the image database and select the image with the highest similarity.

case, an attacker can reveal the identity of the original person.

In this paper, we examine the risk of information leakage in privacy-preserving data augmentation methods that utilize conditioned diffusion models. Recently, an ECCV oral paper [15] introduced an augmentation technique called *instance augmentation* that uses pre-trained diffusion models to redraw sensitive parts of an image. The authors claim that augmented faces can only be identified [5] in 0.14% of all cases for the COCO dataset [18]. Although this is a great achievement, we demonstrate that it is in fact not sufficient to preserve privacy.

We evaluate the following threat model: An attacker seeks to determine the presence of a known individual within an image that has been transformed using an instance-level augmentation technique (*e.g.*, [15]). The attacker does **not** have access to the original image but possesses a set of non-anonymized reference images depicting the target person.

Successfully revealing the individual's identity can have serious consequences. For instance, if the person appears in sensitive datasets - such as those related to protests, medical research, or controversial contexts - exposure could lead to harassment, blackmail, or doxxing. In fields like healthcare or banking, unauthorized access to private medical or financial data could result in fraud or discrimination.

In this paper, we demonstrate that even a simple neural network architecture is sufficient to identify individuals in a face dataset anonymized using instance-level augmentations. To achieve this, we leverage SimCLR [4], a contrastive learning method, to train a model that generates latent representations which encode a person's identity from the input face image. The model is trained to produce similar latent representations for images of the same individual while ensuring distinct representations for different individuals. To identify a person in the dataset, we compute cosine similarity between the latent vectors of the anonymized input image and those of the attacker's image database. A visualization of our method can be found in Fig. 1. Using this approach, we successfully re-identify 69.7% of individuals in our test dataset, which contains 3,421 different people.

We hypothesize that the success of our method stems from the preservation of edges and depth in the instance augmentation process. To test this assumption, we attempt to reveal individuals' identities using only edge images. This scenario can also be viewed as a black-box attack, where the attacker has no access to the analyzed augmentation pipeline [15] or any augmented images. To train our model, we generate edge images using either a HED detector [30] or a Canny edge detector [1], without applying any anonymization. Notably, this approach does not require the attacker to have knowledge of the augmentation pipeline. After training, we apply the model to identify individuals in anonymized images. First, the anonymized images are converted into edge representations using the same detector as in training. Then, we compute similarity scores as in the standard setting. On a dataset containing 3,421 individuals, our black-box method successfully identifies the correct person in an anonymized image with a probability of 4.9%.

To further investigate our hypothesis, we conduct an ablation study on the conditioning of the instance augmentation method, which can preserve edges, depth, or only the segmentation mask. Our experiments show that preserving either edges or depth alone is sufficient for our attack to succeed. Notably, depth preservation inherently retains edge information, as edges are a key component of facial depth. A similarity analysis of edge images across different augmentation variants further supports this hypothesis.

All these findings suggest that even simple methods can effectively re-identify individuals in instance-level augmented datasets if edge information is retained. This raises significant concerns about the use of such techniques for privacy-preserving anonymization.

However, instance-level augmentation remains a powerful technique for improving model performance. It effectively expands datasets with more diverse images while preserving corresponding annotations. Nonetheless, as our work demonstrates, such methods are not suitable for anonymization. The primary issue lies in the conditioning of the diffusion model. Since the model is conditioned to reconstruct the original edges and/or depth map, a neural network can learn this correspondence. This learned information can then be exploited for privacy-revealing attacks.

Our contributions can be summarized as follows:

1. We identify an information leak in data anonymization frameworks that rely on instance-level augmentation techniques using conditioned diffusion models.
2. We demonstrate that contrastive learning is a simple yet highly effective method for identifying individuals in an anonymized dataset.
3. We hypothesize that the preservation of edges is the primary factor driving the success of our method and provide supporting evidence through ablation studies, where we train our model on differently augmented images.
4. We extend our approach by introducing a black-box method that does not require access to the anonymization pipeline yet still successfully identifies individuals.

The code, weights, and generated datasets can be found here: https://lorjul.github.io/compromised-instance-anonymization

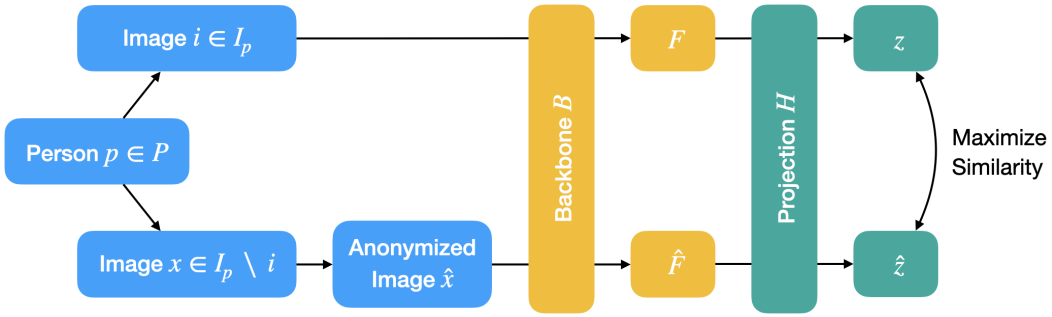

Figure 2. Our contrastive learning framework. I_p is the set of all images from person p. An image i and an augmented image $\hat{x}$, originating from the same person are passed through the backbone B and the projection head H. The similarity of the resulting representations z and $\hat{z}$ is maximized during training.

2. Related Work

2.1. Privacy-Preserving Methods and Attacks

Protecting the privacy for sensitive content in images has been a longstanding challenge in computer vision, with methods including pixelation, blurring [12] or other classical approaches like P3 [27]. However, methods that fundamentally alter the underlying image have a usability tradeoff [10] as parts of the image might not be recognizable anymore. Generative models provide a solution to this problem. Instead of merely destroying information, they can replace parts of the image with generated data [2, 3, 13, 16, 17, 29, 31].

However, improvements on neural network architectures also facilitate attack methods [7]. Prior work has developed methods to defeat classical methods [22], but also more sophisticated methods have been rendered ineffective by deep learning methods [9, 26, 28, 32].

2.2. Instance Augmentation

In this work we re-evaluate the anonymization capabilities of *Dataset Enhancement with Instance-Level Augmentations* [15]. In their work, the authors present a data augmentation method that uses diffusion models to redraw selected instances in an image. To ensure instance replacements that still fit the associated ground truth, they condition the diffusion process using ControlNet [33] and T2I Adapter [24]. These methods ensure that the segmentation mask, depth map, and edge map of the augmented images are similar to their original counterparts. They show that their augmentation strategy improves performance on object detection, semantic segmentation, and salient object detection.

Additionally, the augmentation method is evaluated as a data anonymization tool. To this end, ArcFace [5] is employed to identify peoples' faces in the COCO dataset [18]. On this evaluation, only 0.14% of all faces can be identified. However, we argue that the reported numbers are misleading because evaluating anonymization requires a different benchmark. Instead, a face dataset (*e.g.* [8, 20, 23]) with multiple images of the same person provides a much more realistic evaluation benchmark as we demonstrate in this work.

3. Method

In this section, we describe the contrastive learning framework that we use as our method, which is based on SimCLR [4]. We further describe the three different evaluation protocols that we use, which cover different attacker strengths.

3.1. Contrastive Learning

Our method is a contrastive learning approach similar to SimCLR [4]. In our dataset, we have a set of persons P and sets of original images I_p per person $p \in P$. For each image $i \in I_p$, we use the analyzed augmentation pipeline Φ [15] to create a set of anonymized images $\Phi(i)$. Given a batch size K, we sample persons $S \subset P$ with $|S| = K$ during training. For each $p_k \in S$, we select a random image $i_k \in I_{p_k}$ and an anonymized image $\hat{x}_k \in \Phi(x)$ with $x \in I_{p_k} \setminus i_k$. For all $k \in K$, we feed the batch of images $i_k, \hat{x}_k$ through the backbone B, resulting in features F_k and $\hat{F}_k$. The features F_k and $\hat{F}_k$ are then projected with a small projection head H to a lower-dimensional representation z_k and $\hat{z}_k$. The network is trained to maximize the similarities between the matching representations z_k and $\hat{z}_k$ while pushing apart the similarities of the representations z_k and $\hat{z}_j, k \neq j$ of different persons. SimCLR [4] has shown that such a projection performs better than computing the similarity of the features F_k and $\hat{F}_k$.

Formally, let the similarity function $s(x, y)$ be defined as

$$s(x, y) = \frac{x^T y}{\|x\| \|y\|}. \tag{1}$$

The loss function for a pair of matching representations is then defined as

$$L(z_k, \hat{z}_k) = -\log \frac{\exp(s(z_k, \hat{z}_k)/\tau)}{\sum_{j=1}^{K} \mathbf{1}_{k \neq j} [\exp(s(z_k, z_j)/\tau) + \exp(s(z_k, \hat{z}_j)/\tau)]}, \quad (2)$$

with $\mathbf{1}_{k \neq j}$ evaluating to 1 if $k \neq j$ and to 0 otherwise. τ is a temperature parameter, which either pushes the similarities further apart or pulls them together depending on its value. Fig. 2 visualizes the contrastive learning framework. The usage of a different base image for the augmented version is shown by using the image indices k_1 and k_2 in the figure.

3.2. Evaluation

To evaluate our methods, we use three different evaluation protocols which we call *full reference evaluation*, *single reference evaluation*, and *few reference evaluation*. We will explain all protocols in the following.

Full Reference Evaluation (Full-Ref). In this evaluation scenario, we consider an adversary with access to a database of multiple images per person, excluding the original image that was anonymized. The similarity between the model's representation of the anonymized image and the images in the database is measured. The anonymized image is considered successfully identified if its representation is most similar to an image that depicts the correct individual.

Single Reference Evaluation (Single-Ref). The second scenario is more challenging. Here, the attacker has a database with only one image per individual, excluding the original image that was anonymized. A successful retrieval requires the model to identify the single correct image within the limited gallery.

Few Reference Evaluation (Few-Ref). In the last scenario, the database consists of at most 5 images per person, which is less challenging than the single reference evaluation but more challenging than the full reference evaluation.

The three described protocols evaluate the results for different attacker strengths. The more images per person the attacker has, the easier it is to identify the persons. We report top 1 and top k scores for multiple values of k for all three protocols.

4. Results

4.1. Dataset

We use the publicly available Large-scale CelebFaces Attributes (CelebA) dataset [20] for our experiments. It consists of 202,599 face images from 10,177 individuals. The number of images per individual ranges from 1 to 35.

Augmentations. We apply instance-level augmentations as suggested by [15] to most images of CelebA. We will refer to this technique as *instance augmentation* in the following. We detect the person's bounding box using GroundingDINO [19] and segmentation masks using SAM [14] as suggested by [15]. However, some masks were not correct and some created augmentations did not pass the NSFW filter included in the instance augmentation pipeline. After removing all such images, our dataset consists of 189,582 images. We provide some qualitative examples for images and their augmentations in Fig. 3.

Split	Persons	Original	Augmented
Train	3,654	100,217	154,672
Validation	1,466	20,697	10,368
Test	3,421	48,226	24,542

Table 1. Statistics of our dataset. The table shows the number of individuals, the number of original images, and the number of augmented images for each subset.

Figure 3. Qualitative results for original images (top row) and augmented images (bottom row) from our dataset which is based on CelebA [20].

Data Splits We split the dataset into training, validation, and test set. We ensure that all images from the same person (anonymized or not) are contained in the same split. This ensures that the model does not benefit if it overfits on a person from the training set. Note that even in Full-Ref evaluation, the evaluation is conducted on images from persons that the model did not see at all during training. The different evaluation scenarios only cover the amount of original images that are available per unknown person. Individuals with many available images (augmented and original) are preferably moved to the training set, as more available images from the same individual are beneficial for training. See Tab. 1 for our dataset statistics.

BS	Full-Ref			Few-Ref			Single-Ref		
	top 1	top 5	top 10	top 1	top 5	top 10	top 1	top 10	top 100
16	39.0	58.4	66.2	41.4	60.5	68.1	17.7	37.0	67.5
32	56.9	73.9	79.6	56.4	72.3	78.1	22.7	**42.3**	**70.0**
64	58.5	73.5	78.8	57.2	71.7	76.9	21.9	40.7	68.0
128	**68.1**	**81.2**	**85.3**	**64.5**	**76.9**	**81.1**	**23.9**	41.4	68.5

Table 2. Evaluation scores in % for different batch sizes for all three evaluation protocols mentioned in Sec. 3.2. Depending on the protocol, we evaluate different top k scores.

4.2. Contrastive Learning

We set the temperature constant $\tau = 0.05$ and use an AdamW optimizer with a learning rate of 10^{-4} during training unless mentioned otherwise. Since we want to show that simple models are sufficient to achieve good results in this task, we use a ResNet50 [11] backbone as in the original SimCLR method [4]. We select the best weights according to our validation set and report results on the test set, which consists of 3,421 unique persons with corresponding 24,542 augmented and 48,226 original images (see Tab. 1). Hence, in the Full-Ref evaluation protocol, the attacker's database consists of all 48,226 original images. In the Single-Ref evaluation scenario, the database contains only 3,421 images, which is only approx. 7% of the full database size. For Few-Ref evaluation, the database size is 15,880 images, which is approx. 33% of the full DB size. In Few-Ref, at most 5 images per person are used, but for some persons only less than 5 images are available.

We present the results for our model trained with different batch sizes in Tab. 2. A batch size of 128 achieves the best results for most evaluations. It identifies 68.1% of the persons correctly if provided with the full database. For 81.2% of the examples, the correct person is in the top five outputs. These results are remarkable, since the total amount of persons is over 3k. When limiting the database to a single image per person, the results are worse for all models. However, our best model still identifies 23.9% correctly on first try and 42.3% in the top-10.

Larger batch sizes have a large influence on evaluations with more images. Single-Ref scores vary slightly, Full-Ref scores significantly. Few-Ref evaluation results are pretty close to Full-Ref results, especially for batch size 32. The Single-Ref results have a large gap to the other evaluation scenarios. Hence, an attacker profits from collecting a few images per person, but collecting large amounts of images per person is not necessary to achieve good results.

Although batch size 128 shows the best results for most experiments, we are unable to test even larger batch sizes since they consume large amounts of computational resources. For the same reason, we select batch size 32 for further ablation studies, since it provides a good trade-off between performance and required computational resources. It surpasses batch size 64 in most of our experiments and even batch size 128 in top-10 and top-100 scores for Single-Ref evaluation.

4.3. Ablation Studies

We hypothesize that the edges of individuals in the augmented images play a crucial role in the success of our method. To validate this, we conduct a series of ablation studies to assess the significance of edge preservation. Additionally, we explore the impact of different backbone architectures and temperature values on model performance.

4.3.1. Influence of Conditioning during Augmentation

We want to investigate the importance of edges for our approach. Therefore, we compare the results of our model trained with differently augmented images:

1. Standard instance augmentation using depth and edge modalities
2. Instance augmentation using only depth modality, with **no** edge modality
3. Instance augmentation using only edges modality, with **no** depth modality
4. Instance augmentation using only segmentation masks

Since instance augmentation is computationally expensive, we evaluate these experiments on smaller datasets with approximately 20k images. Each dataset consists of approximately the same number of individuals, as well as augmented and original images. More details can be found in Sec. 7 in the supplementary.

Results. The results are shown in Tab. 3. These scores are not directly comparable to our base model, as we use significantly smaller datasets for training and evaluation. The evaluation dataset consists of approximately 100 individuals and around 3k images. We observe that conditioning the augmentations only on depth or only on edges results in slightly lower performance compared to the standard setting, which includes both conditions. However, the decrease in performance is not as significant as anticipated. Upon visually inspecting the augmented images, we find that depth-only conditioned images preserve the edges almost as well as those conditioned on edges alone or both

(a) original (b) depth and edges (c) depth (d) edges (e) segmentation only

Figure 4. Qualitative examples for augmented images with different conditioning modalities based on CelebA [20] (Top) and corresponding edges detected with a HED detector [30] (Bottom).

conditions. Based on this, we conduct an experiment with images conditioned solely on the segmentation mask. Examples of augmented images for all conditioning variants are shown in Fig. 4. Conditioning only on the segmentation mask leads to the worst results by a large margin, with only 3.2% of individuals correctly identified in the Single-Ref evaluation.

For further edge analysis, we compute the edges of all datasets using a HED detector [30]. Examples of these images are provided in the bottom row of Fig. 4. We then compare the Structural Similarity Index (SSIM) and the mean L1 distance to the edges of the original images. The results, shown in Tab. 4, reveal that the standard augmentation with both conditioning factors produces edges that are most similar to the original, though the difference is minimal compared to conditioning on edges-only or depth-only. This supports our visual observation that depth and edge conditioning preserve edges similarly well. Omitting both conditionings results in images with significantly lower edge similarity.

Depth	Edges	Segm.	Full-Ref		Single-Ref	
			top 1	top 10	top 1	top 10
✓	✓	✓	72.1	91.0	22.4	60.0
✓	-	✓	65.8	87.7	18.1	56.0
-	✓	✓	64.0	86.6	17.9	53.6
-	-	✓	21.7	44.3	3.2	14.9

Table 3. Ablation study on instance augmentation conditioning. We provide experiments for augmented images conditioned on depth and edges, depth only, edges only, and none of both. Conditioning on the segmentation mask (Segm.) is always included. Results include top-1 and top-10 accuracy in % for Full-Ref and Single-Ref evaluation.

Depth	Edges	Segm.	SSIM ↑	L1 ↓
✓	✓	✓	0.62	0.086
✓	-	✓	0.59	0.092
-	✓	✓	0.61	0.094
-	-	✓	0.51	0.128

Table 4. SSIM and mean L1 scores for edge similarity of differently augmented images compared to the original images.

4.3.2. Architecture

Apart from studying the influence of augmentation conditioning, we also conduct ablation studies regarding the temperature parameter of our contrastive learning pipeline. Furthermore, we investigate the usage of other backbone architectures.

Temperature. The temperature parameter (τ) plays a critical role in the similarity calculation. Lower values of τ increase the separation between similarities, making it harder for the model to pull matching pairs closer together. The results, shown in Tab. 5, demonstrate that using $\tau = 1$ yields extremely poor performance, highlighting the importance of this parameter during training. Reduced temperatures result in longer training durations to achieve convergence. Interestingly, our model performs best with lower temperature values, which contrasts with the findings of SimCLR [4]. In their setting, $\tau = 0.1$ achieved the best results, and further decreasing τ led to worse performance.

In our case, the model trained with $\tau = 0.05$ delivers the best scores for the Single-Ref evaluation, while the model trained with $\tau = 0.01$ achieves the best results for the Full-Ref evaluation.

	Full-Ref		Single-Ref	
τ	top 1	top 10	top 1	top 10
1	0.7	5.2	0.8	5.6
0.5	3.3	14.9	2.8	13.4
0.1	33.9	60.5	16.3	36.1
0.05	56.9	79.6	**22.7**	**42.3**
0.01	**61.1**	**80.7**	22.1	39.3

Table 5. Evaluation scores in % for different temperature values.

Backbone. Our main goal is to show that a simple model is capable of identifying persons in instance augmented images of persons' faces. Therefore, we used ResNet50 as a backbone so far, since it is a simple, easy to use backbone. However, we are interested if we can tune the performance of our method further when using a more recent backbone. Hence, we choose ConvNeXt [21] and ViT [6] pretrained with DINOv2 [25] as alternative backbones. We select model variants with a similar amount of parameters compared to ResNet50, resulting in ConvNeXt-tiny and ViT-S/14. We execute experiments with batch size 32 and 128 for both variants and show the results in Tab. 6. The number of parameters per model is also included in the table.

Backbone	Param		Full-Ref		Single-Ref	
		BS	top 1	top 10	top 1	top 10
ResNet	26 M	32	56.9	79.6	22.7	42.3
		128	68.1	85.3	23.9	41.4
ConvNeXt	29 M	32	55.2	78.4	23.1	43.6
		128	**69.7**	**88.6**	26.3	46.1
ViT-S	21 M	32	46.6	72.4	21.2	42.0
		128	53.2	77.1	23.5	44.5
ViT-B	86 M	32	58.0	79.1	26.0	48.7
		64	60.7	82.0	**26.6**	**48.8**

Table 6. Evaluation scores in % for different backbone architectures: ResNet50, ConvNeXt-tiny, ViT-S/14, and ViT-B/14. Note that for all models except ViT-B, we evaluate trainings with batch size 32 and 128, but for ViT-B, the larger batch size is 64.

As expected, modern backbones further improve performance. ConvNeXt with a batch size of 128 outperforms the ResNet50 model by 1.6% in top-1 accuracy for Full-Ref evaluation and by 1.4% for Single-Ref evaluation. ConvNeXt appears to benefit more from a larger batch size, showing a greater improvement over its batch size 32 variant compared to ResNet. For Full-Ref evaluation, ResNet with a batch size of 32 outperforms ConvNeXt. Despite being pretrained with DINOv2, the ViT-S model performs worse than the convolutional models on most evaluation metrics. For a batch size of 128, ViT-S outperforms ResNet in top-10 accuracy, though its top-1 score is slightly lower. One possible reason is that ViT-S has 21 M parameters, 19% less than ResNet50. To further explore the potential of a ViT backbone, we train a ViT-B model with 86 M parameters, over three times the size of ResNet50. While this comparison isn't entirely fair, it provides insight into the ViT's capabilities. As shown in Tab. 6, the larger ViT-B model yields better results. For batch size 32, it outperforms all other models trained with the same batch size. Due to its larger size, we cannot train ViT-B with batch size 128 but include a training with batch size 64 to assess the impact of batch size on performance, resulting in the best scores for Single-Ref evaluation across all models. The results show that a larger batch size generally improves the outcome, even for larger models.

Data Split. In this experiment, we investigate the attacker's strength in a scenario where the attacker's model can be trained using augmented images of the target persons. Hence, we distribute images from the same person across the training, validation, and test set, and call this data splitting variant *distributed*. An attacker who has access to the anonymization pipeline might use this to their advantage and fine-tune a model on a certain person. This experiment emulates this scenario by allowing models to remember people from the training set.

Data Split	Single-Ref		
	top 0.03%	top 0.1%	top 1.0%
Separated	27.9	33.4	56.6
Distributed	22.8	32.0	61.3

Table 7. Evaluation scores in % for different data splits. Either each individual is contained in only one data split (Separated) or the images of each individual are contained in every data split (Distributed).

Results are displayed in Tab. 7. We show top-k scores with a relative k, since the distributed test dataset contains more persons (7,970 to be precise) than the separated test dataset. Interestingly, a model in the distributed setting performs worse for low k, but better for high k. A reason could be that in the distributed setting, all images are distributed across all splits, leading to a significantly lower number of distinct training pairs, which harms the training especially for contrastive learning methods. Moreover, overfitting could play a role, because the model learns highly specific identity-related features that do not generalize well to unseen images of the same person, which is especially relevant for low values of k. However, our method performs

similarly well if it has no access to the images of the target person. Including such images in the training is only beneficial for attacks where a high top-k score for a high k is relevant.

5. Black-Box Attack: Identifying Individuals Using Edge Images Only

The previously discussed methods assume that an attacker has access to anonymized images. We extend our analysis to a black-box attack where an attacker does not require any anonymized images or access to the anonymization pipeline.

The analyzed anonymization strategy uses diffusion models that are conditioned on edges and depth from the original image. Therefore, it is instructed to preserve these edges in the anonymized version. As we already investigate in Sec. 4.3.1, the preservation of the edges is the main reason why our model is successful. To further prove this point, we train another contrastive learning model directly on edge-transformed images that are not anonymized. At test time, anonymized images are transformed to their edge representations and used as input to the trained model for person retrieval. This approach ensures that the model requires no access to the anonymization framework while still allowing for evaluation on anonymized images. As before, we make sure that all images from the same person are either in the training set or in the test set but not in both. We run experiments using Canny edge detection [1] combined with a 5×5 Gaussian blur and using HED [30] to transform the original images to edge images. See Fig. 5 for example edge images.

Figure 5. Edge transformations of the original image (left) using Canny edge detector [1] (middle) and HED [30] (right).

The results are presented in the top half of Tab. 8. Even without any anonymized images during training, the HED-based approach correctly identifies the target individual in the top-10 results with a probability of 4.9%. This is remarkable compared to random guessing, which would yield an accuracy of approximately 0.03%. The Canny-based approach performs significantly worse but remains 20 times more effective than random guessing.

To further investigate the importance of edge preservation, we conduct a similar experiment but allow access to anonymized images during training. In this setting, we train on edge images extracted from augmented images. The results, shown in the bottom half of Tab. 8, indicate that access to augmented images or the augmentation pipeline drastically improves the attacker's success. HED-based edges again yield superior performance, revealing the identity of 50.5% of individuals in the top-10 Full-Ref evaluation. While the absolute score remains lower than training directly on augmented images, these findings demonstrate that edges preserve critical identity-related information, making them a key factor in the success of our method.

Edges	BB	Full-Ref		Single-Ref	
		top 10	top 100	top 10	top 100
Canny	✓	0.6	4.3	0.6	5.8
HED	✓	4.9	18.7	4.6	20.6
Canny	-	24.2	54.9	16.9	42.5
HED	-	50.5	76.6	27.2	54.1

Table 8. Evaluation scores in % for training and evaluating our model on edge images. The top half displays the results for the black-box (BB) attack, the bottom half the results for training on the edges of augmented images.

6. Conclusion

We investigate a recent method, presented as an ECCV oral, that uses conditioned diffusion models as an anonymization strategy. Refuting prior claims, we demonstrate that a simple contrastive learning approach is sufficient to reliably retrieve the anonymized person from a pool of candidates.

We analyze the performance of contrastive learning for person retrieval, finding that a model trained on original and anonymized image pairs is capable to retrieve the correct person out of 3,421 candidates with a probability of 69.7%. We provide a thorough analysis of the trained model and show that the performance depends on the fact that many edges are preserved during the anonymization process. Although the person might seem different to the human eye, a neural network can use the features of the edges to correctly disambiguate between the different people.

Furthermore, we show that preserved image features like edges enable a black-box attack where the attacker does not require any access to anonymized images or the anonymization pipeline. We train a model directly on edge images and show that this is sufficient to retrieve people from a set of candidates.

Using conditioned diffusion models is a great way to increase training data and improve model performance. However, based on our findings, we strongly advise against using methods that preserve features like edges for data anonymization.

References

[1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 2, 8

[2] Jingyi Cao, Bo Liu, Yunqian Wen, Yunhui Zhu, Rong Xie, Li Song, Lin Li, and Yaoyao Yin. Hiding among your neighbors: Face image privacy protection with differential private k-anonymity. In *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6, 2022. 3

[3] Jingyi Cao, Xiangyi Chen, Bo Liu, Ming Ding, Rong Xie, Li Song, Zhu Li, and Wenjun Zhang. Face de-identification: State-of-the-art methods and comparative studies, 2024. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2, 3, 5, 6

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 2, 3

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7

[7] Andrea Gadotti, Luc Rocher, Florimond Houssiau, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*, 10(29):eadn7053, 2024. 3

[8] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 3

[9] Simon Hanisch, Julian Todt, Jose Patino, Nicholas Evans, and Thorsten Strufe. A false sense of privacy: Towards a reliable evaluation methodology for the anonymization of biometric data. *Proceedings on Privacy Enhancing Technologies*, 2024. 3

[10] Rakibul Hasan, Eman Hassan, Yifang Li, Kelly Caine, David J. Crandall, Roberto Hoyle, and Apu Kapadia. Viewer experience of obscuring scene elements in photos to enhance privacy. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Steven Hill, Zhimin Zhou, Lawrence Saul, and Hovav Shacham. On the (in)effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies*, 2016, 2016. 3

[13] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization, 2019. 3

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4

[15] Orest Kupyn and Christian Rupprecht. Dataset enhancement with instance-level augmentations. *arXiv preprint arXiv:2406.08249*, 2024. 1, 2, 3, 4

[16] Tao Li and Min Soo Choi. Deepblur: A simple and effective method for natural image obfuscation, 2021. 3

[17] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 56–65, 2019. 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 3

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3, 4, 6

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[22] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning, 2016. 3

[23] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. Diversity in faces, 2019. 3

[24] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 3

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[26] Kai Packhäuser, Sebastian Gündel, Nicolas Münster, Christopher Syben, Vincent Christlein, and Andreas Maier. Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data. *Scientific Reports*, 12(1):14851, 2022. Publisher: Nature Publishing Group. 3

[27] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. P3: Toward Privacy-Preserving photo sharing. In *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 515–528, Lombard, IL, 2013. USENIX Association. 3

[28] Jimmy Tekli, Bechara Al Bouna, Gilbert Tekli, Raphaël Couturier, and Antoine Charbel. Leveraging deep learning-assisted attacks against image obfuscation via federated

learning. *Neural Computing and Applications*, 36(25): 15667–15684, 2024. 3

[29] Kartik Thakral, Shashikant Prasad, Stuti Aswani, Mayank Vatsa, and Richa Singh. Toonergan: Reinforcing gans for obfuscating automated facial indexing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10875–10884, 2024. 3

[30] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2, 6, 8

[31] Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 5303–5313, New York, NY, USA, 2022. Association for Computing Machinery. 3

[32] Guangsheng Zhang, Bo Liu, Tianqing Zhu, Andi Zhou, and Wanlei Zhou. Visual privacy attacks and defenses in deep learning: a survey. *Artificial Intelligence Review*, 55(6): 4347–4401, 2022. 3

[33] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3

On the Importance of Conditioning for Privacy-Preserving Data Augmentation

Supplementary Material

7. Dataset for Ablation on Conditioning

In Sec. 4.3.1, we compare how well faces can be retrieved when the employed instance augmentation model is not guided by edges or depth maps. Since computing instance-augmented images for the full dataset is computationally expensive, we rely on a smaller dataset for this ablation.

The ablation dataset contains 800 individuals, which are a subset of the CelebA faces that we use for the main experiments. For each individual, we compute on average 25 augmented images.