Fabio Hellmann, Elisabeth André, Mohamed Benouis, Benedikt Buchner, Silvan Mertes*

Anonymization of Faces

Technical and Legal Perspectives

This paper explores face anonymization techniques in the context of the General Data Protection Regulation (GDPR) amidst growing privacy concerns due to the widespread use of personal data in machine learning. We focus on unstructured data, specifically facial data, and discuss two approaches to assessing re-identification risks: the risk-based approach supported by GDPR and the zero or strict approach. Emphasizing a process-oriented perspective, we argue that face anonymization should consider the overall data processing context, including the actors involved and the measures taken, to achieve legally secure anonymization under GDPR's stringent requirements.

1 Introduction

The rapid integration of machine learning into various aspects of our lives, from healthcare to finance and beyond, and the widespread utilization of personal data for training machine learning models have led to growing privacy and security concerns in recent years.

In response to these concerns, legislation such as the European Union's General Data Protection Regulation (GDPR) has been enacted to protect individual privacy and prevent the disclosure of sensitive information. The GDPR has raised awareness of privacy issues and stimulated the development of innovative machine-learning techniques to improve data protection. However, it has become apparent that these techniques can sometimes conflict with other important principles, particularly transparency and accuracy.

While most of the existing technical and legal literature has focused on privacy enhancement methods for structured data, such as tables, there is little research on unstructured data, such as images [1]. This paper focuses on anonymizing facial data as an example of unstructured data. Truong et al. [2] offers a comprehensive examination of privacy-enhancing techniques, particularly evaluating their adherence to the GDPR. Building upon their foundational analysis, our paper explicitly investigates methods for face anonymization, underscoring their importance within the broader array of privacy preservation methods.

Face anonymization obscures personal information such as identity, race, ethnicity, gender, and age, thus minimizing the risk of re-identification. It is vital for sensitive data sets such as medical records and law enforcement data, where maintaining anonymity is paramount. In healthcare, patient confidentiality is ensured when sharing medical images with researchers or healthcare professionals. In criminal justice, anonymizing faces protects the identity of witnesses, victims, and suspects and protects them from potential harm.

According to Weitzenboeck et al.,[1] determining re-identification risks in anonymized data can be assessed through two main approaches: the risk-based approach and the zero or strict approach. GDPR supports the former – it aims to achieve a balance of fulfilling the need for detailed data in research or analysis and the demand for strong privacy protections. It also ensures compliance with legal and ethical requirements while generating accurate and insightful results for academic researchers. In contrast to this approach, which allows for some chance of someone being identified, the zero-tolerance approach prioritizes complete anonymity. However, achieving this level of anonymity often requires deleting the original dataset, which is practically not feasible in most cases.

Given the challenges of achieving absolute anonymity, our paper focuses on a risk-based approach, which involves assessing and mitigating risks to an acceptable level rather than eliminating them. This methodology aligns with GDPR's risk assessment guidelines, advocating for a balanced response to privacy risks. We first discuss the GDPR's Core Principles, stressing the need for clear, legally compliant criteria for face anonymization, and explore techniques like obfuscation, adversarial methods, differential privacy, and latent representations. Our discussion extends beyond the efficacy of these techniques, advocating for a comprehensive view that considers the data's nature, the trustworthiness of involved parties, and the overall data processing context to meet GDPR standards.

* Fabio Hellmann, Elisabeth André, Mohamed Benouis, Silvan Mertes, Chair of Human-Centered Artificial Intelligence; Benedikt Buchner, Chair of Civil Law, Liability Law and Digitalization Law (all University of Augsburg)

[1] Weitzenboeck, E. M., Lison, P., Cyndecka, M., and Langford, M. The GDPR and unstructured data: is anonymization possible? International Data Privacy Law 12, 3 (03 2022), 184–206.

[2] Truong, N., Sun, K., Wang, S., Guitton, F., and Guo, Y. Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Computers Security 110 (2021), 102402.

2 Core Principles of GDPR

Recital 26 of the GDPR clarifies that the principles of data protection law do not apply to anonymous information. Anonymous information is understood, according to Recital 26, as "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." Furthermore, recital 26 also provides a rough guideline for determining whether a natural person is identifiable or not: Account shall be taken "of all the means reasonably likely to be used" and to ascertain whether means are reasonably likely to be used to identify the natural person, "account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

Recital 26 sets out the above-mentioned risk-based approach to classifying information as personal or anonymous. Data is only personal if there is a reasonable risk of identification. If, on the other hand, the risk "appears in reality to be insignificant"[3], the data is considered anonymous, even if identification of the respective person cannot be ruled out with absolute certainty.[4] The purely hypothetical possibility of identifying a person is insufficient to classify this person as identifiable and, therefore, classify the data as personal [5]. However, the GDPR does not contain any more specific requirements that would enable a precise and legally secure differentiation between personal data on the one hand and anonymized data on the other. As a result, classifying data as anonymized is subject to a considerable degree of legal uncertainty.

Take the example of Germany: At the starting point, it is generally accepted that absolute ("perfect") anonymization is neither possible nor necessary under data protection law. In its position paper on anonymization, the Federal Commissioner for Data Protection and Freedom of Information (BfDI) also focuses on so-called de facto anonymization, i.e., it regularly considers it sufficient that "re-identification is not practically feasible because the personal reference can only be restored with a disproportionate amount of time, cost and manpower".[6] However, there is a lack of reliable criteria for when data anonymization can be legally assumed. Instead, further discussion is lost in general discussions on the difficulties of reliable anonymization. The BfDI position paper is ultimately limited to the vague conclusion that "valid anonymization – depending on the type of data to be anonymized and the context of the processing – is a challenge for the respective controller" and, therefore, "sufficient anonymization should not be assumed prematurely".[7]

Against this backdrop, there is an urgent need to develop clear criteria for legally compliant anonymization. The challenges are even greater in the case of facial anonymization, to which a particularly strong personal reference is inherent. However, many anonymization techniques are now available that can contribute to anonymizing this particularly sensitive data. Still, they must be evaluated and categorized to determine whether they can create a reliable basis for classification as anonymous data.

3 Face Anonymization Techniques

This section will present techniques commonly used for anonymizing faces: Obfuscation, Adversarial Techniques, Differential Privacy, and Latent Representations. Remember, these aren't always clearly distinct techniques but somewhat different strategies that can be intertwined.

3.1 Obfuscation

Some methods, called "obfuscation techniques", help protect people's privacy when their faces appear in photos or videos. These methods change or hide certain parts of the face in the image. The goal is to make it hard to recognize the person but keep some general features that don't reveal their identity.

For example, Jourabloo et al.[8] developed a way to hide a person's identity in a photo while keeping important facial features. This method was very good at making it hard to recognize the person. They used a special model and an algorithm to create new images of faces by averaging certain features.

Yang et al.[9] introduced a method that blurs faces in a large dataset of images. Raval et al.[10] used a special mechanism to protect the visual information in video feeds without significantly affecting how the feeds work.

These obfuscation techniques are very effective at protecting privacy, but they make the images look less natural and lower their quality. This limits how these images can be reused for different facial image applications.

3.2 Adversarial Techniques

There are many ways to hide a person's identity in photos, and these methods often use "adversarial techniques". This means that they're trying to do two opposite things at the same time. For hiding identities in photos, these two things are: 1) making sure the person can't be recognized, and 2) keeping features that are important for other tasks.

Trying to do both of these things at the same time is like a tug-of-war game. If you pull too hard on one side (like making sure the person can't be recognized), you might lose something on the other side (like the essential features for other tasks). So, these methods have to find a balance between the two.

For example, Nasr et al.[11] developed a method to minimize the chance of recognizing the person while maximizing the defense

[3] ECJ 19.10.2016 – C-582/14, DuD 2017, 42 – Breyer.

[4] Finck M, Pallas F, They who must not be identified-distinguishing personal from non-personal data under the GDPR, International Data Privacy Law, Volume 10, Issue 1, February 2020, Pages 11-36.

[5] Art. 29 Data Protection Working Party, Personal Data, WP 136 (2007), p. 17.

[6] BfDI. Position paper on anonymization under the GDPR with special consideration of the telecommunications industry. 29.6.2020.

[7] BfDI. Position paper on anonymization under the GDPR with special consideration of the telecommunications industry. 29.06.2020; p. 4.

[8] Jourabloo, A., Yin, X., and Liu, X. Attribute preserved face de-identification. In 2015 International Conference on Biometrics (ICB) (2015), IEEE, pp. 278–285.

[9] Yang, K., Yau, J. H., Fei-Fei, L., Deng, J., and Russakovsky, O. A study of face obfuscation in ImageNet. In International Conference on Machine Learning (2022), PMLR, pp. 25313–25330.

[10] Raval, N., Machanavajjhala, A., and Cox, L. P. Protecting visual secrets using adversarial nets. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017), IEEE, pp. 1329–1332.

[11] Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC conference on computer and communications security (2018), pp. 634–646.

against certain attacks. Another group by Wu et al.[12] used a special model to learn how to change the images to balance recognizing actions with privacy in videos. Yet Wu et al.[13] introduced a method that creates images of faces that can't be recognized but still have important features. They did this by adding certain losses into the training process of their model.

3.3 Differential Privacy

"Differential privacy" is a way to protect privacy that depends on the specific application. In deep learning, which is a type of artificial intelligence, differential privacy involves adding random noise (or random information) to a model that's being trained. This noise is added to ensure the results are balanced, considering both usefulness and privacy.[14] In other words, it's a way to balance making accurate predictions with the model and protecting privacy. For example, Croft et al.[15] could hide identities in images by using differential privacy in a certain way. However, using differential privacy in real-world situations can be challenging. Finding the right balance is essential because adding noise to protect sensitive information might change the data too much, leading to images that can't be recognized.[16]

3.4 Latent Representations

Traditional models used for generating images, known as Generative Adversarial Network (GAN), often have difficulty keeping complex facial features, like emotion, pose, and background. This is because images can be very complex and have a lot of information. This often makes changes in the style of faces less noticeable than changes made directly to the image. "Latent representation" is a simplified version of the data that keeps important features and eliminates unnecessary information. This makes it easier for models to classify and generate images.

For example, Le et al.[17] introduced StyleID, a GAN that changes images into a latent representation, finds essential features, and changes these features. However, StyleID might keep facial traits that could lead to bias or unfairness, even if they're not directly related to identity.

Other methods, such as Sun et al.[18], Hu et al.[19], and Maximov et al.[20], use "inpainting" along with GANs to hide identities in faces based on facial landmarks. These methods are effective, but they keep information relevant to the context outside of the area of the face, like hair color, hairstyle, and gender.

On the other hand, Hukkelås and Lindseth introduced DeepPrivacy2 [21], a more advanced GAN framework for hiding identities in human figures and faces. DeepPrivacy2 uses three detection components for each task: face detection, dense pose estimation, and instance segmentation. They also trained three specific GANs to generate either human figures with conditions, human figures without conditions, or faces.[22] However, using inpainting in these methods might accidentally keep context-relevant information, which could lead to bias or unfairness.

The research by Hellmann et al.[23] introduces a framework for anonymizing faces in images while keeping their emotional expressions intact. This framework employs a GAN to generate an anonymized face version. The unique aspect of this framework is its capability to retain the emotional expressions of the face while eliminating identifiable attributes. This means the anonymized face will not be recognizable as the original individual, but the emotional state displayed by the face, such as happiness or sadness, remains the same. The effectiveness of GANonymization was evaluated in two main areas: increasing anonymity by removing identifiable facial attributes and preserving facial expressions. The results indicated that GANonymization was successful in both areas, effectively anonymizing faces and preserving their emotional expressions. It also demonstrated reliable performance in removing various facial traits, such as jewelry and hair color.

4 Discussing Face Anonymization Techniques in the Light of GDPR

The anonymization procedures listed above clarify that various techniques are now available for facial anonymization. However, 100% anonymity can hardly be guaranteed for faces in particular, at least not if the cognitive value of facial data is not to be lost entirely. It must be assumed that data at an individual level will always contain a last remnant of potential personal reference due to unique combinations of characteristics or other correlations. Even if facial data has been subjected to one of the anonymization procedures listed above, there is always a residual risk of re-identifiability, so the "anonymity" and "re-identification risk identical to zero" can never lead to practicable results.

[12] Wu, Y., Yang, F., Xu, Y., and Ling, H. Privacy-protective-GAN for privacy preserving face de-identification. Journal of Computer Science and Technology 34 (2019), 47–60.

[13] Wu, Z., Wang, Z., Wang, Z., and Jin, H. Towards privacy preserving visual recognition via adversarial training: A pilot study. In Proceedings of the European Conference on Computer Vision (ECCV) (2018), pp. 606–624.

[14] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016), pp. 308-318.

[15] Croft, W. L., Sack, J.-R., and Shi, W. Obfuscation of images via differential privacy: from facial images to general images. Peer-to-Peer Networking and Applications 14 (2021), 1705–1733.

[16] Yoon, J., Drumright, L. N., and Van Der Schaar, M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics 24, 8 (2020), 2378–2388.

[17] Le, M.-H., and Carlsson, N. StyleID: Identity disentanglement for anonymizing faces. arXiv preprint arXiv:2212.13791 (2022).

[18] Sun, Q., Ma, L., Oh, S. J., Gool, L. V., Schiele, B., and Fritz, M. Natural and effective obfuscation by head inpainting. CoRR abs/1711.09001 (2017).

[19] Hu, S., Liu, X., Zhang, Y., Li, M., Zhang, L. Y., Jin, H., and Wu, L. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022), pp. 15014–15023.

[20] Maximov, M., Elezi, I., and Leal-Taixé, L. Ciagan: Conditional identity anonymization generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), pp. 5447–5456.

[21] Hukkelås, H., and Lindseth, F. DeepPrivacy2: Towards realistic full-body anonymization. In Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023), pp. 1329–1338.

[22] Hukkelås, H., Smebye, M., Mester, R., and Lindseth, F. Realistic full-body anonymization with surface-guided GANs. CoRR abs/2201.02193 (2022).

[23] Hellmann, F., Mertes, S., Benouis, M., Hustinx, A., Hsieh, T., Conati, C., Krawitz, P., and André, E. Ganonymization: A GAN-based face anonymization framework for preserving emotional expressions. CoRR abs/2305.02143 (2023).

Against this background, the example of facial data makes it particularly clear that the question of sufficient anonymity of data must not be narrowed down solely to the perspective of whether the data itself has been sufficiently securely anonymized but that the process and the overall context of the data processing must also be taken into account for the question of anonymity. In concrete terms, this means that the question of sufficient anonymization must consider the data properties of the processed faces, the trustworthiness of the actors involved, and the technical and organizational measures taken. It is, therefore, a matter of a process-oriented perspective that takes into account the actual framework conditions and possibilities of identification in the specific data processing procedure.

A process-oriented perspective on anonymity also characterizes the GDPR. In this respect, reference should once again be made to Recital 26 of the GDPR, according to which, to determine whether a natural person is identifiable, all means "reasonably likely to be used" by the controller or another person to directly or indirectly identify a person should be taken into account. The decisive factor under the GDPR is also the identifiability in the specific data processing process. The decisive factor is whether it is possible to assign information to a specific person under the respective framework conditions of data processing with a realistic expenditure of time, costs, and manpower. Based on this understanding of anonymity, the anonymization techniques outlined here, combined with accompanying technical and organizational measures, can then be used to anonymize even such sensitive and unique data as facial data in a legally secure manner.

Acknowledgment

This research was partially funded by the BMBF, Project UBIDENZ (grant number 13GW0568F).

Open Access

Funding

Open Access funding enabled and organized by Projekt DEAL.

Kristian Kersting · Christoph Lampert
Constantin Rothkopf Hrsg.
Wie Maschinen lernen
Künstliche Intelligenz verständlich erklärt
SACHBUCH
Springer

Mario Donick
Die Unschuld der Maschinen
Technikvertrauen in einer smarten Welt
EBOOK INSIDE
Springer