8

Structural Sensitive-Attribute Leakage in Face Recognition Embeddings for Edge AI Deployments

Erica Liu, Enrique Orozco Olivares, Gijs Dubbelman, and Jean-Paul Linnartz

Eindhoven University of Technology, Netherlands

Abstract

Face recognition systems, increasingly deployed on edge devices and servers, store high-dimensional embeddings from models like ArcFace for identity verification and search. While prior work shows such embeddings can reveal sensitive attributes, most target a single attribute at the per-user level in the same-dataset settings. We present a unified study of both per-user and population level demographic leakage, quantifying whether an adversary can predict attributes for individuals and recover aggregate demographic distributions of entire databases. We evaluate both within-dataset and cross-dataset transfer settings. In the within-dataset case, classifiers are trained and tested on the same dataset, while in the cross-dataset case, models trained on one dataset are applied to another, simulating realistic adversaries without access to the target model. Using FairFace and a pretrained ArcFace extractor, we measure leakage of gender, age range, and race, with per-user AUCs exceeding 90% for gender and age range, and over 60% for race (obtained with basic MLP classifiers, making these estimates conservative). Population level estimates recover demographic proportions with low error. We further assess common blurring defences, highlighting recognition–privacy trade-offs. Results indicate demographic leakage is structural, persisting from individual users to entire databases, and across different datasets, underscoring the need for

stronger privacy safeguards in deployed face recognition systems, particularly in edge AI environments.

Keywords: face recognition, privacy risk, population level, cross-dataset transfer.

8.1 Introduction

Edge AI face recognition systems have been increasingly applied in real-world scenarios. The systems normally perform identity verification directly on edge devices by computing high-dimensional embeddings instead of transmitting raw images. These embeddings are typically produced by convolutional neural networks such as ArcFace and are stored on edge devices to efficiently compare and match with the data on central servers. This architecture is often assumed to improve privacy because raw images never leave the device. However, prior work found that embeddings themselves can unintentionally reveal sensitive personal information, making the systems not entirely privacy-preserving as expected. [1, 2].

Previous research has demonstrated that face embeddings can encode attributes such as gender, age, or race [3, 4]. Most analyses focus on inferring a single sensitive attribute at the per-user level within a single dataset or the same model setting, leaving open questions about the generality and structure of such leakage. Recent studies on fairness and privacy risks in biometrics have also emphasized that these representations may encode demographic information even when the original data are anonymized [5, 6].

In this work, we present a systematic framework that unifies per-user level and population level analyses of sensitive attributes leakage of high-dimensional CNN face embeddings. Our approach quantifies not only whether an adversary can infer sensitive attributes for an individual (per-user leakage), but also whether aggregate demographic distributions of an entire embedding database can be reconstructed (population level leakage), enabling large-scale profiling even without certainty at the individual level.

We evaluate leakage under two complementary settings: (i) within-dataset, where attribute probes are trained and tested on the same dataset to measure direct predictability, and (ii) cross-dataset, where probes trained on one dataset (FairFace) are applied to another (UTKFace) to model realistic transfer by adversaries lacking access to the target data. This models an attacker who exploits differences in data distribution—such as varying

demographics or image conditions—while still using the same embedding model.

Using the FairFace and UTKFace datasets and a pretrained ArcFace feature extractor, we measure leakage of gender, age range, and race under varying degrees of Gaussian blurring. Our results demonstrate strong per-user leakage (AUC values larger than 0.9 for gender and age, and larger than 0.6 for race). The low mean absolute error (MAE) between the true proportion and the predicted proportion of one class reveals that population level demographic proportions can still be accurately recovered with low aggregate error. These findings suggest that information leakage in face embeddings is structural, persisting across datasets, aggregation levels, and image preprocessing. This persistence represents an underexplored privacy risk for deployed Edge AI face recognition systems and underscores the need for stronger privacy safeguards in such deployments.

8.2 Related Work

8.2.1 Face Embeddings and Edge AI Deployment

Face recognition models such as FaceNet, CosFace, and ArcFace map facial images to high dimensional embeddings that support efficient matching and storage [7-9]. These embeddings are attractive for the deployments on the edge due to latency and bandwidth constraints, and because raw images can be avoided. However, prior work shows that embeddings and biometric templates may encode more information than only identity.

8.2.2 Sensitive-Attribute Leakage and Template Inversion

A large body of research demonstrates that demographic attributes can be inferred from facial representations, raising fairness and privacy concerns [10]. Beyond attribute inference, template inversion attacks reconstruct faces directly from deep templates, showing that embeddings are not inherently safe [11, 12]. These findings motivate systematic evaluations of what non-identity information is exposed by embeddings under realistic conditions.

8.2.3 Datasets for Attribute Analysis and Bias Measurement

Balanced or demographically annotated datasets such as FairFace and UTK-Face support studying leakage, bias, and generalization [13, 14]. FairFace

improves balance across seven race groups and is widely used for bias measurement, while UTKFace covers a broad age span with age/gender/ethnicity annotations.

8.2.4 Privacy–Utility Trade-offs and Obfuscation

Obfuscation (e.g., blurring, adversarial perturbations) can reduce recognizability, but often exhibits an asymmetric trade-off where utility degrades faster than privacy risk [15]. System-level work further frames obfuscation as a tuneable privacy–utility mechanism across sensing modalities [16]. Our study explicitly quantifies this trade-off by measuring per-user leakage versus recognition utility under Gaussian blurring.

Beyond simple image-space blurring, a rich line of work proposes privacy-enhancing techniques at the image-, template-, and feature levels. Differential-privacy-based protocols perturb face templates or intermediate features, such as PEEP and learnable frequency-domain DP mechanisms for face recognition [17, 18]. Other approaches use adversarial learning to construct privacy-preserving templates: Wang et al. generate adversarial facial features (AdvFace) that defend against feature-to-image reconstruction while preserving recognition accuracy [19]. GAN-based de-identification methods such as Privacy-Protective-GAN [20] and later anonymization frameworks learn to synthesise or modify faces such that identity is removed but task-relevant information is preserved [21, 22].

Recent surveys on privacy-enhancing face biometrics and privacy-preserving face recognition (PPFR) provide a broader taxonomy of such defences across data generation, representation learning, and template storage [6, 23, 24]. Closer to our focus on sensitive attributes in embeddings, template-level regularisation and attribute unlearning methods seek to suppress specific attributes (e.g., gender) while retaining identity. Rezgui et al. enforce angular constraints in the embedding space to reduce gender separability without significantly harming recognition performance [25]. More generally, Guo et al. formalise attribute unlearning as selectively removing information about designated attributes from feature representations via mutual-information-guided detachment losses [26]. These works highlight that attribute factors are geometrically structured within modern embedding spaces and can be actively manipulated.

Our work is complementary to these defences. Instead of proposing a new protection mechanism, we study structural sensitive-attribute leakage that persists in off-the-shelf ArcFace embeddings under a simple and widely

deployed obfuscation technique (Gaussian blurring). By jointly analysing per-user and population level leakage, and by evaluating cross-dataset transfer, we show that demographic information remains recoverable in the embedding space even when (i) per-user inference becomes harder and (ii) image-level utility is severely degraded. This suggests that the leakage quantified in our framework constitutes a baseline risk that privacy-preserving representation learning and attribute unlearning methods must overcome.

8.2.5 Generalization and Threat Modelling

Most prior leakage studies focus on single-attribute, single-dataset, or same-model settings, leaving open how leakage generalizes across data distributions. We contribute a cross-dataset transfer evaluation (train on one dataset, test on another) to model distribution shift in realistic deployments. In parallel, the security community has examined attacks beyond attribute inference, including membership inference on representation-learning systems (e.g., Re-ID) [27], underscoring the need to evaluate embeddings under broader adversarial goals.

8.2.6 Positioning

In contrast to prior work that treats attribute inference or obfuscation in isolation, we provide a unified view that (i) measures per-user leakage and extends to population level aggregates, (ii) evaluates cross-dataset generalization under blur preprocessing, and (iii) ties these findings to a concrete privacy–utility analysis for edge deployments.

8.3 Methodology

We propose a unified framework to quantify sensitive-attribute leakage in deep face embeddings at both the per-user level and population levels.

Given a pretrained face recognition model that maps an input image x to an embedding vector $\mathbf{z} \in R^d$ (in our case d is 512), our goal is to evaluate to what extent $\mathbf{z}$ reveals sensitive attributes such as gender, age, or race.

The evaluation pipeline consists of three stages:

- Embedding extraction: Obtain embeddings $\mathbf{z}$ using a pretrained model (ArcFace) for all face images under varying blurring levels.
- Leakage estimation: Train simple classifiers (MLPs) to predict sensitive attributes from embeddings.

- Privacy quantification: Measure privacy leakage using per-user level and population level metrics, and analyse the trade-off between privacy and utility.

8.3.1 Embedding Extraction and Gaussian Blurring

Let $f_\theta(\cdot)$ denote the feature extractor (We use ArcFace backbone specifically) producing embeddings

$$\mathbf{z} = f_\theta(x) \in R^{512}. \tag{8.1}$$

To simulate privacy-preserving preprocessing, we apply Gaussian blurring to the input images before feature extraction. This technique aims to obscure fine facial details that could aid attribute inference while retaining coarse structure necessary for recognition.

The blurred image $\tilde{x}$ is obtained by convolving the original image x with a Gaussian kernel G_σ:

$$\tilde{x}(u,v) = (x * G_\sigma)(u,v) = \sum_{i=-r}^{r} \sum_{j=-r}^{r} x(u-i, v-j)\, G_\sigma(i,j), \tag{8.2}$$

where

$$G_\sigma(i,j) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{i^2+j^2}{2\sigma^2}\right). \tag{8.3}$$

The kernel radius r is chosen as

$$r \approx 3\sigma, \tag{8.4}$$

which ensures that 99.7% of the Gaussian's total energy is contained within the kernel.

This practical rule balances accuracy and efficiency by cutting off the kernel without significant loss of smoothing effect.

We vary the blur radius $r \in \{0, 1, 2, 3, 5, 8, 10, 12, 15, 20\}$ to analyse how increasing blurring affects both recognition utility and privacy leakage.

The resulting embeddings are then computed as

$$\mathbf{z}_r = f_\theta(\widetilde{x_r}) \tag{8.5}$$

and serve as inputs to the leakage analysis.

8.3.2 Leakage Classifiers

For each sensitive attribute $a \in \{\text{gender}, \text{age}, \text{race}\}$, we train a multilayer perceptron (MLP) g_{ϕ_a} to predict the attribute label from embeddings:

$$(\hat{y})_a = g_{\phi_a}(\mathbf{z}) \tag{8.6}$$

Each MLP consists of an input layer of size 512, two hidden layers (256 and 128 units) with ReLU activations, and a Softmax output layer.

Training minimizes the cross-entropy loss:

$$\mathcal{L}_a = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C_a} y_{i,c}^{(a)} \log\left((\hat{y})_{i,c}^{(a)}\right), \tag{8.7}$$

Where C_a is the number of attribute classes for a.

8.3.3 Metrics for Privacy Leakage

Per-User Leakage: For individual-level leakage, we compute the test accuracy and test area under the ROC curve (AUC) to quantify the classifier's ability to infer sensitive attributes from embeddings.

Higher ACC and AUC indicate stronger attribute predictability and thus higher privacy leakage.

Formally, for classifier scores s_i and ground-truth labels y_i, AUC is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(t)\right) dt, \tag{8.8}$$

where TPR and FPR denote the true positive and false positive rates.

Population Level Leakage: To evaluate aggregate demographic inference, we compute the mean absolute error (MAE) between the true and predicted demographic distributions.

Let $\mathbf{p} = [p_1, \ldots, p_K]$ denote the true proportions of K demographic categories, and $\widehat{\mathbf{p}}$ the predicted proportions from classifier outputs:

$$\text{MAE} = \frac{1}{K}\sum_{k=1}^{K} |p_k - \widehat{p_k}|. \tag{8.9}$$

Lower MAE indicates that the adversary can more accurately reconstruct the population distribution.

For convenience, we report (1 − MAE) as a proxy privacy-leakage score for comparison with utility.

8.3.4 Cross-Dataset Transfer Evaluation

To assess the generalization of leakage across data domains, we perform a cross-dataset evaluation. A classifier trained on embeddings from one dataset, specifically FairFace, is tested on embeddings from another dataset UTKFace. This simulates a realistic adversary who leverages distributional similarities without having access to the exact target-domain data.

8.3.5 Recognition Utility Estimation

Recognition accuracy of the underlying embedding model is used as a measure of system utility. However, datasets such as FairFace and UTKFace lack consistent identity labels, making direct identity verification or retrieval evaluation infeasible. Given this limitation and the fact that ArcFace verifies embeddings using cosine similarity, we estimate proxy utility as the average cosine similarity between embeddings of original (non-blurred) and blurred images:

$$\cos\left(\mathbf{z}_0, \mathbf{z}_r\right) = \frac{\mathbf{z}_0 \cdot \mathbf{z}_r}{||\mathbf{z}_0||\ ||\mathbf{z}_r||} \tag{8.10}$$

As cosine similarity serves as the standard metric in ArcFace-based recognition systems and reflects representational separability, the trade-off between privacy and utility is analysed by comparing ACC, AUC and (1 − MAE) against the average cosine similarity across blur levels.

8.4 Experiments

8.4.1 Experimental Setup

All experiments are conducted using Python 3.8 and PyTorch 2.1 on an NVIDIA RTX 4060 GPU. Face embeddings are extracted using the pretrained ArcFace model (ResNet50 backbone) from the InsightFace repository. For each experiment, we use the same embedding extractor without fine-tuning to ensure that observed leakage arises from representational properties rather than retraining.

Classifier Training: For each sensitive attribute (gender, age, race), an MLP classifier g_{ϕ_a} is trained on top of the fixed embeddings as described in methodology. We use a 70/15/15 split for training, validation and testing, and

repeat experiments with five random seeds to report mean results. Training employs the Adam optimizer with learning rate 10^{-3}, batch size 128, and early stopping based on validation loss.

Blur Levels: Gaussian blur radius $r \in \{0, 1, 2, 3, 5, 8, 10, 12, 15, 20\}$ are applied to each image. A radius of $r = 0$ corresponds to the unblurred baseline. The corresponding standard deviation follows $\sigma \approx r/3$. Each blur level generates a new embedding set, allowing us to jointly evaluate privacy leakage and recognition utility across multiple privacy intensities.

8.4.2 Datasets

We use two publicly available face datasets with balanced demographic annotations:

- **FairFace:** contains $108,501$ face images labelled for gender, age range, and race across seven demographic groups. It is designed to provide balanced representation across races and age ranges, making it ideal for measuring leakage in unbiased embeddings.
- **UTKFace:** includes approximately $23,700$ cropped faces annotated with age, gender, and ethnicity. It features diverse lighting, pose, and background conditions, supporting generalization evaluation.

We conduct both within-dataset and cross-dataset evaluations:

- **Within-dataset:** classifiers are trained and tested on the same dataset.
- **Cross-dataset:** classifiers are trained on FairFace and evaluated on UTKFace, simulating an adversary trained on a different but related data distribution.

Additionally, for consistency across datasets, we harmonize race categories by mapping FairFace's seven race labels to the five used in UTKFace. For age prediction, we convert the continuous or multi-class age labels into binary groups using cutoffs at 18, 30, and 50, which simplifies the task.

8.4.3 Evaluation Protocols

Per-User Leakage: For each attribute, the MLP predicts probabilities $\widehat{y_a}$ over attribute classes. The per-user privacy leakage is quantified by the AUC between predicted scores and true labels. We report the mean and standard deviation of AUC across five runs for each blur level. Higher AUC indicates stronger attribute inference capability and therefore higher leakage.

Population Level Leakage: To quantify aggregate demographic inference, we compute the predicted demographic proportions $\widehat{\mathbf{p}}$ from classifier outputs and compare them with ground-truth proportions $\mathbf{p}$ using the MAE defined in equation (8.9).

Recognition Utility: Recognition utility is estimated via the average cosine similarity between embeddings at different blur levels, as defined in equation (8.10). Cosine similarity serves as a proxy measure of representational quality: a higher average cosine similarity implies that embeddings preserve more discriminative structure. We analyse the privacy–utility trade-off by jointly examining ACC, AUC (per-user leakage), (1 − MAE) (population leakage), and cosine similarity (utility) across blur levels.

8.4.4 Results and Analysis

8.4.4.1 Per-User Leakage Results

Figure 8.1 shows test ACC values for gender, age, and race inference across blur levels. Even with moderate blurring ($r = 5$), classifiers achieve ACC > 0.85 for gender and most age ranges, and around 0.6 for race. The test AUC results for age range and gender are shown in Figure 8.2. Similarly, the classifier can achieve AUC > 0.80 even at higher blur ($r = 8$). These

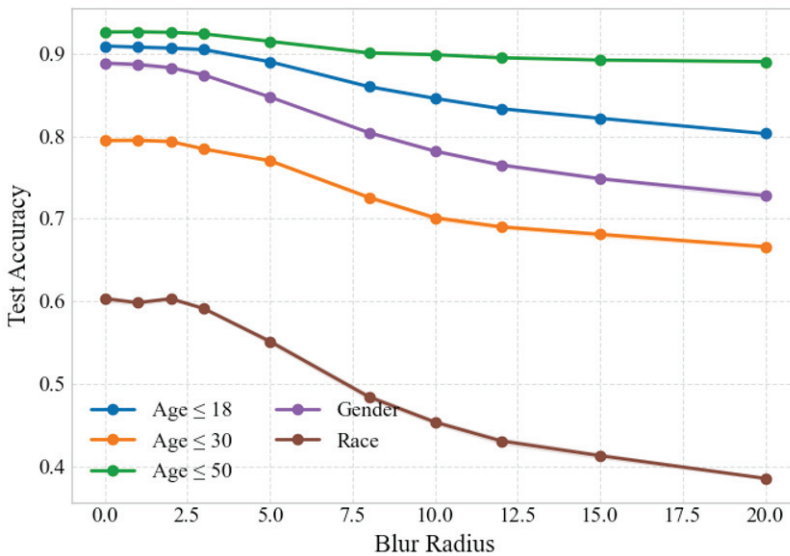

Figure 8.1 Test ACC across blur levels by age, gender, and race on FairFace.

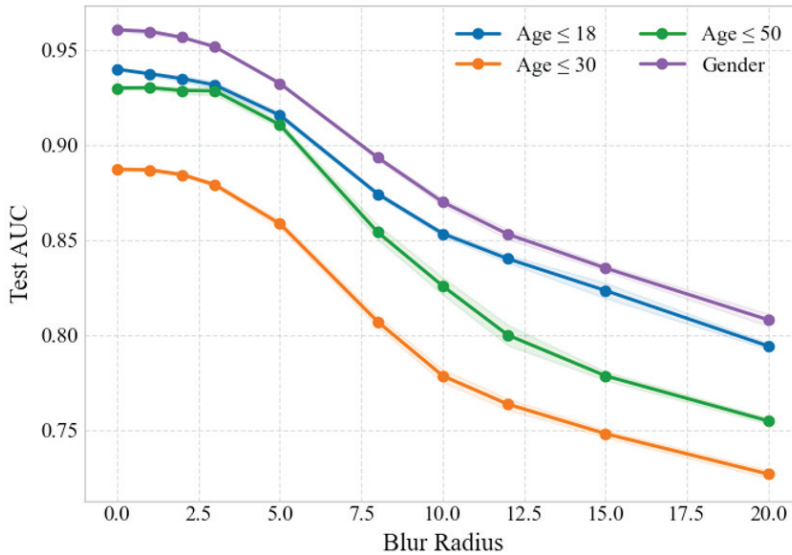

Figure 8.2 Test AUC across blur levels by age, gender on FairFace.

indicate that sensitive attributes remain highly predictable even when visual details are suppressed. The reduction in ACC and AUC with stronger blurring is modest, suggesting that demographic information is embedded structurally in the feature space.

8.4.4.2 Population Level Leakage

Table 8.1 presents the MAE scores measuring population level reconstruction accuracy, in which the lower MAE means higher leakage. Although the MAE values fluctuate without a clear monotonic trend, this variation is likely due to the overall small magnitude of errors. Results reveal that aggregate demographic distributions can be recovered with extremely low error (MAE $< 30\times\ 10^{-3}$) even when blur level $r\ = 20$, where per-user sensitive attribute leakage has already dropped substantially. This implies that an adversary can estimate the demographic composition of an entire embedding database even without confidently predicting each individual's attributes.

8.4.4.3 Cross-Dataset Generalization

In the cross-dataset setting, attribute classifiers retain substantial predictive power, achieving ACC$\approx\ 0.8$ for gender and age range on unseen domains as shown in Figure 8.3. Blurring provides limited privacy protection in this

Table 8.1 Population level privacy leakage based on FairFace embeddings (MAE $\times$ 10^3; mean $\pm$ std) across blur levels for age, gender, and race.

Target Label	**Blur Radius (r)**									
	0	**1**	**2**	**3**	**5**	**8**	**10**	**12**	**15**	**20**
Age 18 cutoff	7.4±1.0	6.4±5.7	4.0±2.2	6.7±3.3	3.9±2.8	3.0±2.0	33.1±12.0	10.4±12.8	23.8±4.3	16.2±4.4
Age 30 cutoff	7.2±4.4	17.2±4.9	14.4±9.4	10.9±7.5	14.4±10.4	9.2±4.9	8.3±8.9	8.1±6.8	8.0±3.7	11.1±10.9
Age 50 cutoff	11.2±1.5	19.8±7.6	23.3±5.9	18.5±10.5	19.0±10.3	11.8±12.7	14.6±12.6	24.3±6.5	14.8±17.1	7.4±8.9
Gender	5.3±4.6	9.4±4.4	14.7±6.2	14.9±2.6	9.3±9.7	14.6±8.6	6.9±4.2	16.9±3.0	18.7±13.1	11.0±8.2
Race	5.6±0.9	5.7±1.2	8.6±3.5	8.3±2.5	7.7±3.4	6.5±1.3	6.6±1.8	8.7±2.5	7.8±0.9	5.8±1.9

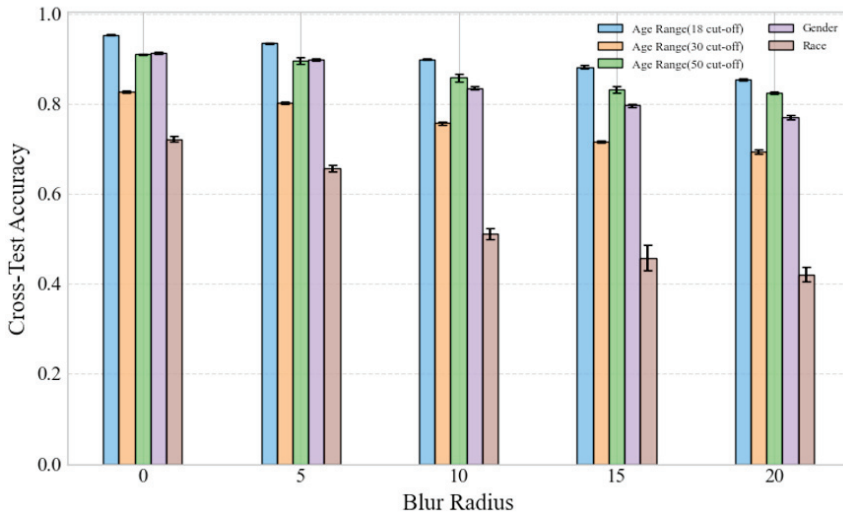

Figure 8.3 Test accuracy across blur levels by age, gender, and race on UTKFace.

Table 8.2 Population level privacy leakage based on UTKFace embeddings (MAE $\times$ 10^3; mean $\pm$ std) across blur levels for age, gender, and race.

Target Label	**Blur Radius (*r*)**				
	0	**5**	**10**	**15**	**20**
Age 18 cutoff	36.3±6.4	34.6±7.4	39.3±22.8	29.1±4.2	40.7±4.4
Age 30 cutoff	16.4±5.3	14.5±7.3	15.8±8.6	34.9±7.8	45.0±17.3
Age 50 cutoff	24.7±14.7	45.9±19.5	67.1±17.5	88.1±2.1	79.1±11.5
Gender	12.7±9.1	11.2±11.1	28.7±7.3	30.6±11.5	35.6±10.5
Race	61.8±6.3	80.8±5.1	110.5±6.7	119.5±7.3	123.9±6.9

setting. Although race test accuracy is lower, this is reasonable given the larger number of classes. This demonstrates that sensitive-attribute leakage generalizes across datasets and is not purely dataset-specific. In addition, the results of population level leakage in the cross-dataset setting are shown in table II. The highest MAE is only around 120 $\times$ 10^{-3}, indicating that even in cross-dataset scenarios with high blur levels, a considerable amount of demographic population level information remains recoverable.

8.4.4.4 Privacy–Utility Trade-Off

Figure 8.4 and Figure 8.5 illustrate the privacy–utility trade-off on UTKFace embeddings, showing how recognition utility (cosine similarity) relates to

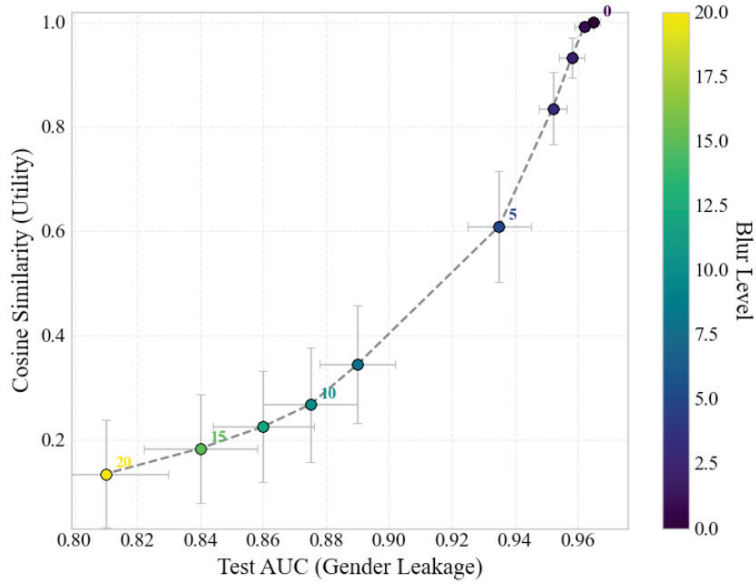

Figure 8.4 The trade-off between utility and per-user level privacy leakage for gender on UTKFace.

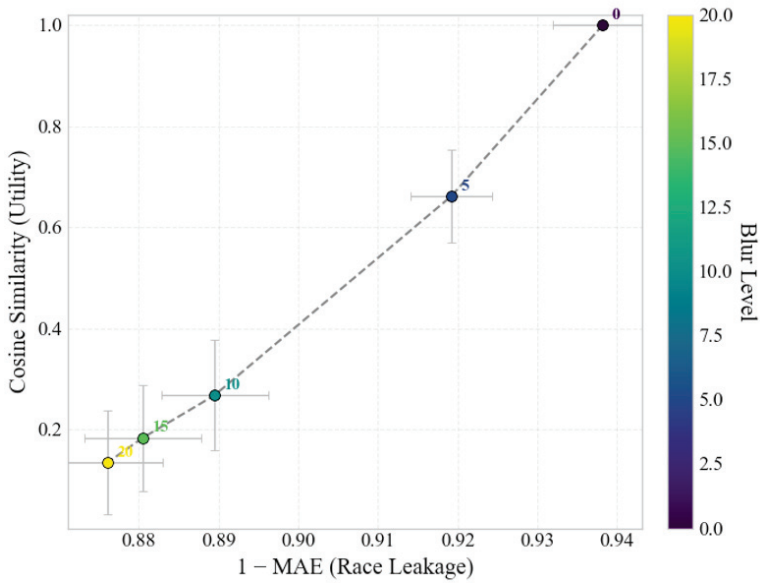

Figure 8.5 The trade-off between utility and population level privacy leakage for race on UTKFace.

privacy leakage metrics (AUC for gender and 1 – MAE for race) across different blur levels. As the blur radius increases, cosine similarity steadily decreases, indicating a degradation in recognition quality. Meanwhile, both gender AUC and 1 – MAE for race show only modest declines, suggesting that blurring provides limited reduction in privacy leakage. Overall, the reduction in leakage is relatively minor compared to the significant loss in utility, revealing an asymmetric trade-off: blurring rapidly impairs recognition performance before achieving meaningful suppression of demographic attribute leakage.

8.4.5 Discussion

To highlight the robustness of structural leakage, we compare per-user and population level metrics at the strongest blur setting ($r = 20$). As expected, heavy Gaussian blurring reduces per-user attribute inference: gender and age AUCs drop noticeably, and race prediction becomes less reliable.

However, population level leakage remains largely unaffected. The MAE between predicted and true demographic proportions stays within a low range, indicating that aggregate demographic structure is still preserved even when individual predictions become uncertain. This suggests that blur disrupts fine-grained cues required for per-user inference but does not eliminate the global geometric patterns in the embedding space that reflect demographic composition.

This asymmetry has important privacy implications: while blurring may appear to protect individuals, it does not prevent reliable profiling of entire embedding databases. Combined with the strong impact of blur on recognition utility, these findings show that image-space obfuscation is insufficient and highlight the need for embedding-level privacy mechanisms.

8.5 Conclusion

This work presented a systematic analysis of sensitive-attribute leakage in deep face embeddings, focusing on Edge AI recognition systems that operate without transmitting raw images. We developed a unified framework to evaluate privacy leakage at both the per-user and population levels, and extended it to cross-dataset settings to assess generalization under distribution shifts.

Experiments on FairFace and UTKFace show that sensitive demographic information remains highly inferable from embeddings even under moderate Gaussian blurring. Per-user inference achieves AUC values above 0.9 for gender and age and around 0.6 for race, while population level demographic distributions can still be reconstructed with low error (MAE $< 50\times\ 10^{-3}$). Notably, population level leakage is often more persistent than individual leakage, as aggregate demographic proportions remain accurately recoverable even when per-user predictability declines. This indicates that demographic leakage is a structural property of modern face embeddings rather than an artifact of overfitting or dataset bias.

The privacy–utility analysis further reveals an asymmetric trade-off: blurring severely degrades recognition utility while offering only limited privacy gains. These results underscore the insufficiency of image-space obfuscation and the need for embedding-level privacy mechanisms.

Overall, our findings demonstrate that privacy vulnerabilities in face embeddings persist across individuals, populations, and datasets, with population level leakage emerging as a particularly robust and underappreciated threat. While our study provides a unified analysis of per-user and population level leakage under blur-based obfuscation, it also has several limitations. In particular, our evaluation is restricted to a single embedding model (ArcFace) and a single image-space defence (Gaussian blurring), which, although widely used in practice, do not cover the full spectrum of modern recognition architectures or privacy-preserving mechanisms. These limitations suggest that structural demographic leakage may vary across models or be mitigated differently by embedding-level defences such as adversarial unlearning or differential privacy. Extending the framework to additional architectures and protection strategies therefore remains a valuable direction for future work.

Future work should explore privacy-preserving embedding transformations that jointly suppress per-user and population level leakage without sacrificing recognition accuracy, including differential privacy, adversarial training, and attribute disentanglement directly in the embedding space.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101097300.

References

[1] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond Identity: What Information Is Stored in Biometric Face Templates?" in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), 2020, pp. 1–10.

[2] C. Song and A. Raghunathan, "Information Leakage in Embedding Models," in Proc. ACM Conf. Computer and Communications Security (CCS), 2020, pp. 377–390.

[3] H. J. Ryu, H. Adam, and M. Mitchell, "InclusiveFaceNet: Improving Face Attribute Detection With Race and Gender Diversity," in Proc. ICML Workshop on Fairness, Accountability, and Transparency, 2018.

[4] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, "Fairness Properties of Face Recognition and Obfuscation Systems," in Proc. USENIX Security Symp., 2023, pp. 731–748.

[5] I. Fábián, "A Comparative Study on the Privacy Risks of Face Recognition Libraries," Acta Cybernetica, vol. 25, no. 3, pp. 233–256, 2021.

[6] L. Laishram et al., "Toward a Privacy-Preserving Face Recognition System," ACM Comput. Surv., 2025, to be published.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 815–823.

[8] H. Wang et al., "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 5265–5274.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 4690–4699.

[10] S. Gong, X. Zhu, and S. Gong, "Jointly De-biasing Face Recognition and Demographic Attribute Estimation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 330–347.

[11] K. Kärkkäinen and J. Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2021, pp. 1548–1558.

[12] Z. Zhang, "UTKFace: Large-Scale Face Dataset," 2017. [Online]. Available: https://susanqq.github.io/UTKFace/

[13] V. Chandrasekaran et al., "Face-Off: Adversarial Face Obfuscation," Proc. Privacy Enhancing Technol. (PoPETs), vol. 2021, no. 3, pp. 356–375, 2021.

[14] N. Raval et al., "Olympus: Sensor Privacy Through Utility-Aware Obfuscation," Proc. Privacy Enhancing Technol. (PoPETs), vol. 2019, no. 1, pp. 5–25, 2019.

[15] M. A. P. Chamikara, P. Bertók, I. Khalil, D. Liu, and S. Camtepe, "Privacy Preserving Face Recognition Utilizing Differential Privacy," Comput. Secur., vol. 97, p. 101951, 2020.

[16] J. Ji et al., "Privacy-Preserving Face Recognition With Learnable Privacy Budgets in Frequency Domain," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022, pp. 475–491.

[17] Z. Wang et al., "Privacy-Preserving Adversarial Facial Features," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 8009–8018.

[18] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-Protective-GAN for Privacy Preserving Face De-Identification," J. Comput. Sci. Technol., vol. 34, no. 1, pp. 47–60, 2019.

[19] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to Anonymize Faces for Privacy Preserving Action Detection," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 620–636.

[20] H. Tian, T. Zhu, and W. Zhou, "Fairness and Privacy Preservation for Facial Images: GAN-Based Methods," Comput. Secur., vol. 122, p. 102902, 2022.

[21] B. Meden et al., "Privacy-Enhancing Face Biometrics: A Comprehensive Survey," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 4147–4183, 2021.

[22] Z. Sun and Z. Liu, "Ensuring Privacy in Face Recognition: A Survey on Data Generation, Inference and Storage," SN Appl. Sci., vol. 7, no. 441, 2025.

[23] Z. Rezgui, N. Strisciuglio, and R. N. J. Veldhuis, "Gender Privacy Angular Constraints for Face Recognition," IEEE Trans. Biometrics Behav. Identity Sci., vol. 6, no. 3, pp. 352–363, 2024.

[24] T. Guo, S. Guo, J. Zhang, W. Xu, and J. Wang, "Efficient Attribute Unlearning: Towards Selective Removal of Input Attributes from Feature Representations," in Proc. ACM Int. Conf. Multimedia (MM), 2022.

[25] J. Gao et al., "Similarity Distribution Based Membership Inference Attack Against Person Re-Identification," in Proc. AAAI Conf. Artif. Intell. (AAAI), 2023, pp. 436–444.
[26] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain, "On the Reconstruction of Face Images From Deep Face Templates," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 5, pp. 1188–1202, 2019.
[27] H. O. Shahreza, M. Rabiee, and H. K. Ekenel, "Face Reconstruction From Facial Templates by Learning Latent Space Mapping," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.