Explainability-Driven Image Anonymization in Latent Space (EDIALS)

Younas Khan
younas.khan@urv.cat
Universitat Rovira i Virgili
Tarragona, Spain

Anna Monreale
anna.monreale@unipi.it
University of Pisa
Pisa, Italy

Carlo Metta
carlo.metta@isti.cnr.it
University of Pisa
Pisa, Italy

David Sánchez
david.sanchez@urv.cat
Universitat Rovira i Virgili
Tarragona, Spain

Josep Domingo-Ferrer
josep.domingo@urv.cat
Universitat Rovira i Virgili
Tarragona, Catalonia

Abstract

Facial image anonymization is essential to enable privacy-preserving image data sharing. The core challenge lies in removing identity-revealing information without degrading the utility of the images, which is essential for, e.g., demographic analysis. However, existing techniques apply uniform pixel-level distortions or synthesize replacements using Generative Adversarial Networks (GANs), which do not retain the meaningful features necessary for downstream tasks. To address this issue, we introduce EDIALS (Explainability-Driven Image Anonymization in Latent Space), which selectively modifies identity-specific latent features identified via explainability techniques. By applying targeted, incremental distortions in the latent space of an adversarial autoencoder, EDIALS effectively anonymizes images while preserving their analytical utility much better than existing techniques. Empirical evaluations on a common dataset show that EDIALS achieves 0.42% re-identification risk (equivalent to random guessing) while maintaining high utility: 84.66% F_1 for age, 97.91% for gender, and 82.61% for race classification. In contrast, DeepPrivacy2 —a state-of-the-art GAN-based approach— results in a re-identification risk as large as 16.27% and lower utility: 78.32% F_1, 82.58%, and 73.32% for age, gender, and race classification, respectively.

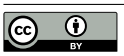
CCS Concepts

• **Security and privacy → Data anonymization and sanitization**; • **Computing methodologies → Machine learning approaches**.

Keywords

Machine Learning, Deep Learning, Face Recognition, Privacy Protection, Utility Preservation

ACM Reference Format:
Younas Khan, Anna Monreale, Carlo Metta, David Sánchez, and Josep Domingo-Ferrer. 2026. Explainability-Driven Image Anonymization in Latent Space (EDIALS). In *Proceedings of the Sixteenth ACM Conference on Data and Application Security and Privacy (CODASPY '26), June 23–25, 2026, Frankfurt am Main, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3800506.3803492



This work is licensed under a Creative Commons Attribution 4.0 International License.
CODASPY '26, Frankfurt am Main, Germany
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2562-3/2026/06
https://doi.org/10.1145/3800506.3803492

1 Introduction

In recent years, there has been an unprecedented surge in image sharing on the Internet, driven primarily by the rapid growth of social media platforms. In 2024, internet users uploaded approximately 3.2 billion images onto various social media platforms daily, Facebook and Instagram alone contributing 825 million and 275 million uploads per day, respectively [9]. Because images containing facial content generate significantly higher engagement (they received 38% more likes and 32% more comments than non-facial posts [2]), users are especially motivated to share facial images, which are the most privacy-sensitive type of images. In fact, human faces contain distinctive identifying features that modern facial recognition (FR) systems can exploit with remarkable precision. Current state-of-the-art FR systems achieve accuracy rates exceeding 99.8% on standard benchmarks [11], enabling nearly perfect re-identification capabilities. This high accuracy transforms casual image sharing into a significant privacy vulnerability, as individuals can be identified and tracked without their knowledge or consent in multiple contexts.

On the other hand, facial images provide more useful and less harmful value beyond personal identification. Many research tasks benefit from facial data analysis without the need to identify specific individuals. For example, medical research uses facial features to study genetic conditions [7, 20], psychological studies analyze emotional expressions [37, 53], and demographic analysis examines age and gender characteristics [30, 46] –all of which do not require identified information.

The core challenge in facial image anonymization is balancing the trade-off between privacy protection and analytical utility preservation. Traditional image anonymization approaches like pixelation and blurring indiscriminately modify all (facial) pixels [39] to prevent re-identification, but also destroy valuable non-identifying information. This uniform distortion does not consider that re-identification is based on specific facial regions and features [50]. Thus, rather than applying uniform distortion, an effective anonymization approach should selectively modify only those features most responsible for re-identification.

In this paper, we propose EDIALS (Explainability-Driven Image Anonymization in Latent Space), a novel approach that selectively modifies only those features that contribute to re-identification. Our method leverages the latent feature space of an adversarial autoencoder (AAE), where facial attributes are represented in a structured and organized manner. By using explainability techniques to identify the specific latent features most relevant to identification, we can consequently apply targeted distortions that prevent recognition while minimizing the impact on other facial/image characteristics.

Contribution. This paper makes the following contributions:

- We introduce EDIALS, an explainability-driven image anonymization approach that leverages the ABELE (Adversarial Black box Explainer generating Latent Exemplars) [18] framework to target identity-revealing features.
- We introduce a novel feature distortion mechanism that selectively modifies identity-revealing features in the latent space based on explainability-derived decision rules.
- We demonstrate through extensive evaluation that our approach achieves a better privacy-utility trade-off than both conventional anonymization methods and state-of-the-art GAN-based techniques.

The remainder of this paper is organized as follows. Section 2 reviews existing image anonymization techniques. Section 3 describes the EDIALS methodology. Section 4 presents our experimental setup, evaluation framework, and the analysis and discussion of the empirical results. Section 5 reports the conclusions and depicts some lines of future research.

2 Related work

The first and most widely adopted facial anonymization techniques are pixelation and blurring [39, 54]. Pixelation reduces image detail by segmenting an image into $n \times n$ blocks and averaging pixel values within each block, where increasing block size improves privacy protection but decreases image quality. Blurring applies Gaussian filters with kernel size k to reduce image sharpness by averaging pixel values with their neighbors, giving more weight to pixels closer to the center of the kernel [38]. Motion blurring is a variant that creates directional distortion effects that simulate movement during image capture [49]. Despite their widespread use, these techniques suffer from a critical limitation: they uniformly modify all image pixels regardless of their contribution to identity disclosure, and this unnecessarily degrades the analytical utility of the images.

More recently, there have been some attempts to enforce differential privacy (DP) on images. Differential Privacy-Pixelation (DP-Pix) applies Laplacian noise to pixelated images [14], while Differential Privacy-Blur (DP-Blur) extends this approach with an additional smoothing step using Gaussian blur [15]. These methods calibrate noise based on the global sensitivity of the output to changes in individual data, theoretically ensuring that the presence or absence of any individual's data in a dataset remains indistinguishable up to an exponential factor ϵ [12]. However, guaranteeing DP on image releases faces significant challenges. For facial images, a meaningful implementation of the DP principles requires noise calibrated to the maximum possible change to any individual's image. This means that most (if not all) pixels might be replaced with maximum/minimum RGB values in a face image, thereby rendering DP-protected images unusable. To retain image utility, practical implementations of DP typically limit noise to a small subset of m pixels, but this compromises the theoretical guarantees that DP aims to provide.

More targeted approaches focus on specific facial features rather than modifying entire images. Eye and mouth masking techniques superimpose black circles or boxes on specific facial regions to emulate sunglasses or face masks [16]. However, their effectiveness varies substantially based on the size of the drawn circles or boxes [45]. Moreover, some studies indicate that masking these regions alone may be insufficient, as FR also relies on other features such as eyebrows [47]. Furthermore, determining common parameters for detecting and masking eyes and mouths in large and heterogeneous datasets remains challenging due to variations in facial structures, poses, and feature sizes.

Recent advances in machine learning have led to Generative Adversarial Network (GAN)-based anonymization methods that generate privacy-preserving synthetic images [17]. StyleGAN-2 [26] employs latent vectors as replacements for original images to generate new ones, whereas Perceptual Indistinguishability-Net (PI-Net) [8] employs an encoder-decoder architecture to produce obfuscated images. More sophisticated methods like Conditional Identity Anonymization GAN (CIAGAN) [35] and DeepPrivacy [25] leverage conditional GANs (cGANs) for image anonymization while preserving background information. DeepPrivacy2 [24] enhances the original DeepPrivacy method by ensuring that the generated face does not appear in the dataset, and by extending the anonymization to the entire body. Recent work has also investigated simplified and generative anonymization pipelines [13, 28], which rely on full-face synthesis and do not explicitly detect or perturb identity-revealing features. While these methods effectively prevent re-identification, they suffer from several limitations: they tend to produce obviously synthetic faces, their code and trained models are often unavailable, and most importantly, they may inadvertently alter characteristics important for demographic analysis when generating entirely synthetic replacements. Explainability-driven anonymization has also been explored in a recent work [22], which motivates selectively distorting identity-relevant features rather than applying uniform distortions.

3 Explainability-Driven Image Anonymization in Latent Space (EDIALS)

A common limitation across existing image anonymization methods is their inability to selectively target identity-revealing features. Conventional methods modify all pixels uniformly, while GAN-based approaches replace entire facial regions with synthetic alternatives. Both strategies unnecessarily compromise features that are valuable for analytical tasks. Research indicates that both humans and algorithms focus on specific facial features for re-identification, such as the eyes and forehead [50]. This insight suggests that effective anonymization should target only those features that contribute significantly to identification while preserving other characteristics.

Table 1: Notation used in EDIALS, including the latent representations, explainability components, and distortion parameters used throughout Section 3

Symbol	Description
x	Input image
z	Latent representation of x
$\mathcal{E}, \mathcal{D}$	Encoder and Decoder of the adversarial autoencoder
b	Black-box model (face recognition classifier)
H	Set of latent vectors generated by *neighgen*
$\tilde{H}$	Valid latent vectors retained after discriminator filtering
δ	Discriminator for latent vector validity
c	Decision tree approximating b
r	Decision rule extracted from c
Z_r	Set of identity-revealing latent features
α	Distortion factor
$Z_{\text{distorted}}$	Distorted latent representation during anonymization
$X_{\text{distorted}}$	Image obtained by decoding $z_{\text{distorted}}$
X_{anon}	Anonymized dataset

This is precisely the cornerstone of our method, EDIALS, which selectively targets identity-revealing features while preserving characteristics essential for analytical tasks such as age, gender, and race classification. In this way, EDIALS enables researchers to train models on anonymized images that can later generalize well to non-anonymized images in real-world applications.

3.1 Image explainer

Our work builds on the foundational framework introduced in the ABELE explainer [18]. ABELE offers a model-agnostic approach for explaining black-box decisions for image classification by leveraging the latent feature space learned through an AAE. The key strength of ABELE lies in its ability to identify features that are deemed important by any black-box model (b) without requiring access to its internal structure. This capability is particularly valuable for our anonymization task, as it enables us to precisely target the features that contribute most significantly to identity recognition. We leverage ABELE's methodology for encoding images into the latent space (z) and extracting decision rules. We then modify it to a distortion-based mechanism tailored to protect face images.

We chose ABELE over explainers like LIME (Local Interpretable Model-agnostic Explanations) [44] because LIME fails to account for pixel relationships, which produces explanations that uniformly cover most of the image and makes them unsuitable for targeted anonymization. ABELE, by contrast, leverages the latent feature space to generate meaningful exemplars that respect feature relationships. Empirical evidence by [18] shows that ABELE outperforms LIME in fidelity (mimicking black-box behavior), coherence, and stability, making it ideal for precisely identifying identity-revealing features.

3.2 Adapting ABELE to Facial Recognition

Since FR represents a significantly more complex task than the general image classification problems addressed in the original ABELE implementation, we have adapted its architecture components while preserving ABELE's core model-agnostic methodology.

For the AAE, we employ a truncated ConvNeXt-Base model [34] pre-trained on ImageNet as the encoder ($\mathcal{E}$), removing the classification layers to capture rich visual features at multiple scales. The encoder outputs a 1024-dimensional bottleneck representation, compressing the image information while preserving semantic details.

The decoder ($\mathcal{D}$) uses a progressive upsampling architecture featuring five transposed convolutional stages that gradually increase the spatial resolution from 8×8 to 256×256 pixels. Each upsampling stage incorporates a transposed convolution followed by LayerNorm [1] and GELU activation [23] for stable training. The final layer employs a tanh activation function to generate normalized RGB images within the [-1,1] range.

The discriminator (δ) consists of convolutional layers that process the latent vectors, followed by fully connected layers that determine whether a latent representation belongs to the prior distribution. To ensure stable adversarial training, we employ hinge loss [31] with gradient penalty regularization [19]. In our setting, the discriminator is used to filter latent vectors by retaining only those satisfying $\delta(h) \geq 0.5$; further details of the implementation are provided in Appendix A.

Together, these components enable high-fidelity reconstruction of facial images from their latent representations while maintaining the semantic details that are essential for both identity protection and utility preservation. Our method employs AAEs rather than conventional autoencoders because they combine GANs with autoencoder learning, ensuring that generated samples follow the same distribution as original images. The AAE architecture allows us to work in a lower-dimensional latent space ($k \ll n$), where features are more structured and semantically meaningful, though not completely disentangled. While perfect feature disentanglement remains an active research challenge, as demonstrated by recent works [36, 52], the AAE's latent space provides sufficient organization of semantic concepts to allow targeted manipulation of identity-related features.

3.3 EDIALS design

Thanks to ABELE's model-agnostic nature, EDIALS is designed with a modular architecture that decouples the encoding-decoding mechanisms from the core anonymization logic. This enables seamless integration of future advancements in representation learning: as more sophisticated autoencoder variants emerge with improved feature structuring capabilities, they can be seamlessly incorporated into our algorithm without redesigning the fundamental anonymization procedure. Similarly, for the black-box model (b) that EDIALS aims to anonymize against, we utilize a ConvNeXt-Base model fine-tuned for FR. This choice reflects the practical reality of modern FR systems, ensuring evaluation against a realistic adversary.

The EDIALS workflow is illustrated in Figure 1, which shows the progress from initial image encoding to final anonymization, and it is formalized in Algorithm 1. It begins with initialization (line 1) of an empty set to store anonymized images (X_{anon}). For each image in the original dataset (X) (line 2), the algorithm encodes

it into its latent representation using the encoder $\mathcal{E}$ (line 3). This transformation is crucial as it enables manipulation in a lower-dimensional space where semantic features are more distinctly represented.

Following ABELE's approach, EDIALS generates a neighborhood of latent vectors (H) around the original encoding through the *neighgen* module (line 4). This neighborhood includes instances with both decisions, i.e., $H = H_{=} \cup H_{\neq}$ where instances $h \in H_{=}$ are such that $b(\tilde{h}) = b(x)$, and $h \in H_{\neq}$ are such that $b(\tilde{h}) \neq b(x)$. This balanced neighborhood is crucial as it enables the decision tree to accurately learn the local behavior of the black-box model.

To ensure that the generated latent vectors correspond to plausible images, the algorithm filters them using δ (line 5), retaining only valid vectors ($\tilde{H}$) for further processing. These valid latent vectors are then decoded back into the image space using $\mathcal{D}$ (line 6), producing a set of images ($\tilde{X}$) that represent valid variations of the original. By querying the black-box model with the decoded images (line 7), EDIALS obtains their classifications (Y) which, combined with the corresponding latent vectors, serve as training data for a decision tree (c) (line 8). This process enables the decision tree to learn how the black-box model makes identity classification decisions in the latent space. The resulting interpretable decision tree approximates the decision boundary of the black-box model, allowing EDIALS to identify which specific latent features most strongly influence identity recognition. From this tree, the algorithm extracts a decision rule (r) and identifies the set of features $Z_r = \{z_1, z_2, \ldots, z_m\}$ that significantly influence the classification outcome (line 9). These identity-revealing features become the target for the image distortion process.

At this point, EDIALS leverages ABELE's explainability insights by implementing a novel, incremental distortion strategy that iteratively perturbs the identity-revealing latent features, applying only the minimal distortion required to prevent re-identification. While ABELE generates both factual and counter-factual rules to explain black-box decisions through exemplars and counter-exemplars, EDIALS deliberately focuses only on the factual rules. This design choice is motivated by our anonymization objective: factual rules directly capture the features that most strongly support the current identity classification, making them ideal targets for privacy-preserving distortion. Counter-factual rules, while valuable for explanations, could lead to unpredictable identity transitions rather than targeted anonymization. By concentrating solely on factual rules, we achieve more controlled and smaller distortions. The core distortion phase initializes a copy of the original latent vector ($z_{\text{distorted}}$) and establishes a feature counter (i) to track the feature currently being modified (line 10). The algorithm then enters an iterative distortion loop that continues until the black-box model misclassifies (i.e., fails to re-identify) the distorted image (line 12), indicating successful anonymization.

Within each iteration, the algorithm modifies one feature at a time by adding or subtracting α (line 13). The direction of the distortion is determined by the condition in the decision rule: if a rule condition specifies that $z_i < threshold$, then α is added to increase the value and violate this condition; conversely, if a rule condition states $z_i \geq threshold$, then α is subtracted to push the value below the threshold. This targeted approach ensures that the distortion directly counters the specific conditions that lead to the correct identity classification. After each feature modification, the distorted latent vector is decoded to produce an image ($x_{\text{distorted}}$) (line 14), which is evaluated by the black-box model to check if misclassification has occurred. The distortion process systematically progresses through the identified features by incrementing the feature counter after each modification (line 15). When all features have been processed once without achieving misclassification, the counter is reset (lines 16-17), allowing multiple passes through the feature set with cumulative distortions.

This targeted and incremental approach ensures minimal image distortion by modifying only the features necessary for preventing re-identification while preserving characteristics contributing to utility-oriented tasks. The key advantage of incremental distortion is that it adapts to each feature's varying sensitivity to perturbation. By testing after each increment, we apply precisely the necessary distortion per feature, avoiding over-distortion. Once an image is successfully misclassified, it is added to the anonymized dataset (line 20). The algorithm continues this process for all images in the original dataset and ultimately returns the complete anonymized collection (line 22).

4 Empirical results

We next report a comprehensive evaluation of EDIALS and compare it to conventional anonymization methods and state-of-the-art approaches. Our analysis focuses on both privacy protection and utility preservation metrics. All experiments were conducted on an AMD Ryzen 5 3600 CPU (3.6 GHz base and max speed) with 32 GB of RAM and an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM). Our implementation uses PyTorch 2.2.2, and the complete code, and datasets are available at https://github.com/yk0803/EDIALS. Additional implementation details for reproducibility are provided in Appendix A.

4.1 Evaluation dataset and metrics

We selected the Multi-Task Face (MTF) dataset [21], which consists of 5,246 images depicting 240 distinct individuals. We preferred the MTF dataset over other commonly used datasets, such as CelebA [33], because the latter contains low resolution images with significant labeling inaccuracies caused by automatic annotation [32]. MTF, on the other hand, provides 1024×1024 resolution images with manually verified labels for four distinct tasks: identity recognition, age classification, gender recognition, and race classification.

This curated and multifaceted annotation makes MTF the best suited dataset for our evaluation, as it enables simultaneous assessment of both privacy protection (through identity recognition) and utility preservation (through the other three classification tasks).

The dataset contains 240 unique identities with a well-balanced distribution across demographic categories: 190 young and 50 old individuals; 130 males and 110 females; and diverse racial representation, including 80 Asian (Chinese and Korean), 49 Asian (Indian), 35 Black, and 76 White individuals. For our experimental protocol, we divided the dataset into 70% for training and 30% for validation and testing across all four tasks.

To evaluate privacy protection, we measure the Re-Identification Risk (RIR) as the ratio of correctly identified images to the total

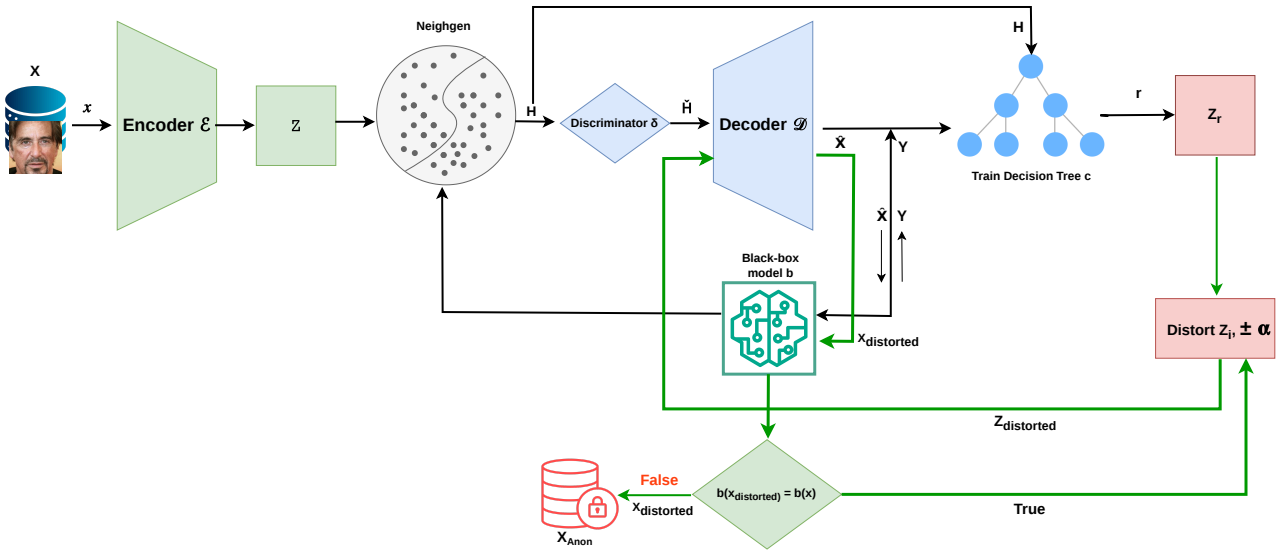


Figure 1: EDIALS workflow: Input x from dataset X is encoded by $\mathcal{E}$ into latent representation z, then processed through neighgen to create neighborhood H. After validation by δ and decoding by $\mathcal{D}$, a decision tree c extracts rule r and features Z_r for targeted distortion. The green pathways illustrate the EDIALS-specific distortion process, where $z_{\text{distorted}}$ is incrementally and iteratively perturbed by α and decoded until b misclassifies (i.e., fails to re-identify) the resulting $x_{\text{distorted}}$, which is then added to X_{anon}.

number of images, bounded below by the random guessing probability:

$$\text{RIR} = \max\left\{\frac{\text{Correct Predictions}}{\text{Total Predictions}}, \frac{1}{C}\right\} \tag{1}$$

where C is the number of identities in the dataset. Since an adversary can always randomly guess among C identities with probability $\frac{1}{C}$, this represents the theoretical minimum RIR. For our dataset with $C = 240$ identities, the minimum RIR is $\frac{1}{240} \approx 0.42\%$.

Predictions are provided by an FR model trained on the original images and identity labels. This simulates a scenario in which an adversary sees clear identified images of the individuals in the dataset. The FR model was implemented with a ConvNeXt-Base architecture [34] pre-trained on ImageNet and fine-tuned for FR by adapting the final classification layer to the number of identities in our dataset. We used cross-entropy loss as the optimization criterion and Adam optimizer [27] with a learning rate of 0.00001. Early stopping was enforced if the model did not improve the validation loss for five consecutive iterations to prevent overfitting.

To assess utility preservation, we trained separate ConvNeXt-Base models for age, gender, and race classification on the anonymized datasets (with identity labels removed). The goal was to check the usefulness of anonymized images for training classification models. We then quantified utility preservation using the F_1 score, calculated as the harmonic mean of precision and recall:

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall.}} \tag{2}$$

The F_1 score provides a balanced measure by considering both false positives and false negatives, making it particularly suitable for evaluating classification performance on the demographic tasks.

4.2 Conventional Methods and EDIALS Results

We begin our empirical evaluation by analyzing six conventional image anonymization techniques: pixelation, Gaussian blurring, motion blurring, facial feature masking, and DP-enhanced variants of pixelation and blurring (DP-Pix and DP-Blur). These methods serve as baselines for further comparison.

We initially experimented with the recommended parameters for each technique: block sizes $n \in [4, 10]$ for pixelation and kernel size $k = 25$ for Gaussian blurring [49]. However, our empirical testing revealed that these values resulted in negligible reductions in RIR. Since our objective was to reduce RIR to levels that approach random guessing while maintaining utility metrics well above baseline random performance (50% for binary classifications such as age and gender, and 25% for the four-class race classification), we expanded the parameter ranges.

Algorithm 1 EDIALS

Require: Dataset X, Encoder $\mathcal{E}$, Decoder $\mathcal{D}$, Discriminator δ, Black-box model b, Distortion factor α
Ensure: Anonymized dataset X_{anon}

Initialize $X_{\text{anon}} \leftarrow \emptyset$
for each image $x \in X$ **do**
  $z \leftarrow \mathcal{E}(x)$
  $H \leftarrow \text{neighgen}(z, b)$
  $\tilde{H} \leftarrow \{h \in H \mid \delta(h) = \text{valid}\}$
  $\tilde{X} \leftarrow \{\mathcal{D}(\tilde{h}) \mid \tilde{h} \in \tilde{H}\}$
  $Y \leftarrow \{b(\tilde{x}) \mid \tilde{x} \in \tilde{X}\}$
  $c \leftarrow \text{DecisionTree.}Train(H, Y)$
  Extract rule r & features $Z_r = \{z_1, z_2, \ldots, z_m\} \leftarrow c$
  $z_{\text{distorted}} \leftarrow z$, $i \leftarrow 1$
  $x_{\text{distorted}} \leftarrow \mathcal{D}(z_{\text{distorted}})$
  while $b(x_{\text{distorted}}) = b(x)$ **do**
    $z_{\text{distorted}}[z_i] \leftarrow z_{\text{distorted}}[z_i] \pm \alpha$
    $x_{\text{distorted}} \leftarrow \mathcal{D}(z_{\text{distorted}})$
    Increment i
    if $i > m$ **then**
      Reset $i \leftarrow 1$
    end if
  end while
  $X_{\text{anon}} \leftarrow X_{\text{anon}} \cup \{x_{\text{distorted}}\}$
end for
return X_{anon}

For DP-based methods, we adopted 'safe' $\epsilon \in \{0.1, 0.5, 1\}$, but the small neighborhood of $m = 16$ employed in [15], with an additional kernel size (k) of 99 for DP-Blur. It is worth noting that $m = 16$ accounts for only 0.0015% of pixels in our 1024×1024 pixel images. Given that a substantial proportion of pixels in facial images contribute to identification, such a small set of pixels cannot provide meaningful DP guarantees. Our attempts to set larger and more theoretically sound values for m (ideally covering the entire image) resulted in completely randomized outputs even with extremely large ϵ values, rendering the images useless for any analytical task. This highlights a fundamental limitation in applying DP to image data: when no *ex ante* privacy guarantees hold due to weak privacy parameters, the actual privacy protection can only be evaluated empirically [3].

As shown in Table 2, increasing the block size in pixelation improves privacy protection, but reduces utility across all classification tasks. This behavior highlights the inherent privacy-utility trade-off: stronger anonymization reduces RIR at the expense of degrading utility, whereas configurations that preserve utility exhibit higher RIR. The race classification task, which requires distinguishing between four different classes, suffers the most significant degradation, dropping from 87.33% at $n = 20$ (and very large RIR) to just 18.25% at $n = 50$ (with reasonably low RIR). This occurs because pixelation obscures distinctive facial features such as eye shape, nose structure, and skin tone variations that are essential for recognizing different racial characteristics. Age classification performance also declines substantially as pixelation removes crucial age-related markers, particularly affecting the ability to identify older individuals who form the minority class in our dataset. Gender classification remains more resilient to pixelation compared to the other tasks, likely because gender recognition can leverage broader facial proportions and contours that remain partially discernible even at lower resolutions.

Gaussian blurring follows a different pattern: while it preserves some of the general facial structure, it significantly degrades fine details and texture information. Age classification is particularly affected by blurring, as this task relies heavily on facial wrinkles, skin texture, and other subtle age-related features that are smoothed out even by moderate levels of Gaussian blur. Race classification also suffers, as blurring homogenizes facial features crucial for distinguishing between racial groups. Gender classification maintains better performance under blurring because it can still utilize the preserved facial contours and proportions that remain visible after the blurring process.

Motion blurring creates directional distortion that has a unique impact pattern. At moderate levels, it disrupts identifying features while maintaining the overall facial structure, achieving better privacy-utility trade-offs than equivalent levels of Gaussian blurring. However, at high intensities (with $k = 300$), motion blur severely distorts all facial features, causing substantial degradation across all classification tasks. Unlike pixelation and Gaussian blur, extreme motion blur particularly impacts gender classification because the directional smearing distorts the overall facial shape and proportions that gender classification typically relies upon, dropping below a random guess (41.67%) at $k = 300$.

We also evaluated masking techniques by simulating the effect of wearing face masks and/or sunglasses. Using Haar-cascade [48] for detection via the OpenCV library [5], we applied masks to hide specific facial features. As shown in Table 2, mouth masking alone proved least effective, with a RIR as large as 94.12%, because mouth regions generally lack strong identity-revealing features. However, utility metrics remained high with only minimal degradation from the original values. Eye masking achieved better privacy protection with RIR reduced to 30.21%, while combining both eye and mouth masking further reduced RIR to 27.27%. The eye masking approach had a moderate impact on utility, particularly affecting race classification (dropping to 61.25%) since eye shape is a key feature for racial identification. Combined eye and mouth masking had more pronounced effects on all utility metrics, though gender classification remained relatively robust at 83.11%.

DP-based methods achieve stronger privacy protection than previous methods, but at a higher cost to the utility, as depicted in Table 2. DP-Pix and DP-Blur both add random RGB noise to the images, which not only obscures identifying features but also introduces strong artifacts that do not allow classification models to generalize to unseen data. At $\epsilon = 0.1$, both methods achieve RIR values close to 1%, indicating near-perfect anonymization. DP-Blur shows slightly better utility preservation than DP-Pix at comparable privacy levels, particularly for gender classification (44.03% vs. 31.68% at $\epsilon = 0.1$). This occurs because the additional blurring step in DP-Blur creates a smoothing effect that mitigates some of the disruptive noise patterns introduced by the DP mechanism. However, at these high privacy levels, utility metrics for both methods approach random guessing, especially for race classification.

EDIALS results: Our EDIALS approach was evaluated using the methodology described in Section 3. For the neighborhood

generation phase, we ensured adequate coverage of the latent space surrounding each image encoding by generating a balanced set of samples representing both positive and negative classifications. This strategy allowed our explainability component to identify decision boundaries with sufficient precision while maintaining computational efficiency. For the distortion component, we applied a balanced factor of $\alpha = 0.5$ to the identity-revealing features identified in the latent space (see Section 4.5 for additional results with other distortion factors). Because EDIALS requires both the AAE and black-box models to remain on the GPU, our hardware resources only allowed us to run EDIALS with reasonable runtimes on images with a maximum resolution of 256×256 pixels. Therefore, MTF images were downscaled to that resolution (see Section 4.6 for an analysis of the influence of image resolution on performance and runtime).

As shown in Table 2, EDIALS achieved the theoretically higher levels of anonymization with RIR of 0.42%, equivalent to random guessing. Simultaneously, it largely preserved utility across all classification tasks, with F_1 scores of 84.66% for age classification, 97.91% for gender classification, and 82.61% for race classification.

The effectiveness of EDIALS comes from its targeted distortion in the latent feature space. By operating in this transformed space rather than directly on pixel values, EDIALS can exploit the structural organization of facial attributes that occurs in well-trained autoencoders. Our explainability component identifies which specific latent dimensions most strongly influence identity recognition, allowing us to modify only those features while leaving other characteristics largely intact.

This selective modification is key to preserving utility. For instance, gender classification remains highly accurate (97.91%) because the latent features that influence gender information appear to be sufficiently distinct from those that strongly influence identity recognition. The explainability-guided approach ensures these gender-defining features remain undisturbed during the anonymization process. Similarly, age-related and race-related features are substantially preserved, allowing for effective demographic analysis even after complete anonymization.

The ability to maintain high F_1 scores for race classification (82.61%) is particularly noteworthy, as this task typically requires the preservation of subtle facial features that conventional methods often destroy when achieving strong privacy protection. By targeting only identity-revealing components in the latent space, EDIALS preserves these critical features that contribute to race classification tasks, such as certain aspects of facial structure and skin tone.

These results demonstrate that the conventional trade-off between privacy and utility can be substantially mitigated through our targeted approach to feature distortion in the latent space.

4.3 Comparison with the state-of-the-art

GAN-based approaches represent the current state-of-the-art in image anonymization. Among these methods, DeepPrivacy2 [24] is a recent and widely adopted approach with publicly available implementation and pre-trained models. We used the implementation provided by the authors[1], which includes a pre-trained GAN model with 46 million parameters.

Our evaluation revealed that DeepPrivacy2 achieved a RIR of 16.27%, indicating weaker than desired privacy protection. To enable a fair comparison at similar RIR levels, we configured EDIALS with $\alpha = 0.5$ and set an iteration limit of 60 (as 68 iterations achieved a RIR of 0.42%). With this configuration, EDIALS achieved a RIR of 15.68%, closely matching DeepPrivacy2's privacy level.

At these comparable privacy levels, EDIALS significantly outperforms DeepPrivacy2 across all utility metrics, as shown in Table 3: age classification (89.32% vs. 78.32%), gender classification (97.98% vs. 82.58%), and race classification (87.14% vs. 73.32%). This performance advantage demonstrates that our targeted approach to identity disruption better preserves features essential for demographic analysis.

The synthetic nature of GAN-generated images introduces artifacts that impair classification models. While DeepPrivacy2's images may appear more detailed than conventional anonymization methods, they exhibit unnaturally synthetic characteristics that affect classification accuracy. Additionally, DeepPrivacy2's complete replacement of facial features can unintentionally alter demographic characteristics essential for utility tasks. EDIALS avoids these issues through its more selective modification approach that targets identity-revealing features while minimizing distortion to characteristics important for demographic analysis.

4.4 Runtime analysis

Computational efficiency is an important consideration when processing large image collections. Table 4 presents the average per-image processing time for each evaluated method.

Simpler methods like pixelation and blurring offer the fastest processing at 0.04 seconds per image. DP-based approaches require slightly more computation due to noise calculation overhead, with DP-Pix and DP-Blur taking 0.13 and 0.17 seconds per image, respectively.

DeepPrivacy2 requires 2.31 seconds per image for inference. This figure excludes the training time, which the authors did not report, but would be substantial given their training data scale (more than 1 million images) and model complexity (46 million parameters).

EDIALS incurs a higher processing time at 256×256 resolution, requiring 33.00 seconds (on average) per image. However, reducing the resolution to 128×128 enables batch processing and reduces the runtime to 9.0 seconds per image, representing a practical operating point. The time reported also includes the complete pipeline: AAE training, black-box model training, encoding, neighborhood generation, decision tree learning, feature distortion, and decoding. The variant that matches DeepPrivacy2's RIR requires 2 fewer seconds as it takes fewer iterations (60 vs. 68) due to fewer decode-evaluate cycles during incremental distortion.

The AAE and black-box models must both remain on the GPU for the technique to work properly, which precludes batch processing with the current hardware. This technical requirement is intrinsic to the operation of the method. However, the runtime of EDIALS decreases substantially with lower-resolution images (Section 4.6), where reduced memory requirements enable batch processing and

[1] Github URL of DeepPrivacy2: https://github.com/hukkelas/deep_privacy2

Table 2: Privacy protection and utility preservation comparison across conventional anonymization methods and EDIALS on the MTF dataset. RIR measures privacy protection, where lower values indicate stronger anonymization, and 0.42% represents the lower bound (random guessing among the 240 identities). Utility is assessed through F_1 scores on three demographic classification tasks: age (binary: young/old), gender (binary: male/female), and race (four-class: Asian Chinese/Korean, Asian Indian, Black, White). Parameter values (n for pixelation block size, k for blur kernel size, ϵ for differential privacy, and α for EDIALS distortion factor) configure varying intensity levels for each technique and balance the privacy-utility trade-off.

Technique	Parameter	Values in %			
		RIR	F_1Age	F_1Gender	F_1Race
Original images	–	100.00	93.09	98.88	94.65
Pixelation	$n = 20$	84.00	96.95	97.43	87.33
	$n = 35$	13.87	53.31	54.25	34.59
	$n = 50$	3.01	47.35	49.17	18.25
Blurring	$k = 50$	77.08	78.19	96.95	82.84
	$k = 100$	13.02	63.30	74.71	29.67
	$k = 200$	2.50	36.81	54.57	23.48
Motion Blur	$k = 100$	61.74	75.85	95.60	79.74
	$k = 200$	19.25	34.77	52.35	42.14
	$k = 300$	7.95	30.29	41.67	26.11
DP-Pix	$\epsilon = 1$	9.83	65.81	91.46	58.51
	$\epsilon = 0.5$	6.34	52.93	44.94	18.16
	$\epsilon = 0.1$	1.09	34.77	31.68	23.55
DP-Blur	$\epsilon = 1$	24.03	68.97	91.47	49.17
	$\epsilon = 0.5$	9.85	42.13	61.62	16.19
	$\epsilon = 0.1$	1.01	29.12	44.03	13.69
Mouth Masked	–	96.12	90.11	95.28	91.65
Eyes Masked	–	30.21	80.13	85.81	61.25
Eyes & Mouth Masked	–	27.27	76.15	83.11	59.15
EDIALS	$\alpha = 0.5$	**0.42**	**84.66**	**97.91**	**82.61**

Table 3: Utility preservation evaluation for EDIALS and DeepPrivacy2 at comparable privacy levels (RIR ≈ 16%), enabling direct comparison of their utility preservation capabilities across the demographic classification tasks. EDIALS was configured with $\alpha = 0.5$ and limited to 60 iterations to match DeepPrivacy2's privacy level, while DeepPrivacy2 results were obtained using the authors' pre-trained GAN model. Original image performance is included as an upper-bound reference.

Method	F_1 Score (%)			
	Age	Gender	Race	RIR (%)
Original images	93.09	98.88	94.65	100.00
DeepPrivacy2	78.32	82.58	73.32	16.27
EDIALS	**89.32**	**97.98**	**87.14**	**15.68**

Table 4: Average per-image processing time comparison. EDIALS includes full pipeline (training + processing); DeepPrivacy2 is inference-only (training time was not reported by the authors).

Method	Runtime (seconds)
Pixelation	0.04
Blurring	0.04
DP-Pix	0.13
DP-Blur	0.17
DeepPrivacy2	2.31
EDIALS	**33.0**

yield processing times comparable to those of DeepPrivacy2 while achieving a superior privacy-utility balance.

4.5 Impact of the distortion factor

This section analyzes how the distortion factor employed at each iteration (α) affects computational efficiency and utility preservation in our approach. Table 5 presents experimental results for different α values.

Table 5: Impact of the distortion factor α on computational efficiency and utility preservation at a fixed privacy level (RIR = 0.42%). Smaller increments require more iterations but yield better utility across demographic tasks. The selected $\alpha = 0.5$ (highlighted) balances efficiency and utility, as further reductions provide diminishing returns at substantially higher computational cost.

Distortion Factor α	Average Iterations	Theoretical Total Distortion*	F_1 Score (%) Age	Gender	Race
1.0	36	3.60	82.22	95.09	80.23
0.7	51	3.57	83.15	96.22	81.32
0.6	59	3.54	83.89	97.05	81.86
0.5	**68**	**3.40**	**84.66**	**97.91**	**82.61**
0.2	168	3.36	85.78	97.91	83.75
0.1	328	3.28	85.80	97.92	83.76

*Theoretical Total Distortion calculated as $(iterations \times \alpha)/10$, where 10 is the average number of features.

All α values achieved anonymization at the level of random guessing (0.42% RIR). As expected, smaller α values required more iterations to reach the same privacy threshold, increasing from 36 iterations at $\alpha = 1.0$ to 328 iterations at $\alpha = 0.1$.

The results show that as α decreases from 1.0 to 0.1, utility improves across all tasks. Gender classification F_1 scores increase from 95.09% to 97.92%, age classification from 82.22% to 85.80%, and race classification from 80.23% to 83.76%. This improvement aligns with decreased total distortion, which drops from 3.60 at $\alpha = 1.0$ to 3.28 at $\alpha = 0.1$.

As introduced above, we used $\alpha = 0.5$ in the main experiments as it provides a good balance between utility and efficiency. While utility barely increases for $\alpha < 0.5$, the computational cost significantly increases. At $\alpha = 0.5$, we achieve F_1 scores of 84.66% (age), 97.91% (gender), and 82.61% (race) with only 68 iterations, compared to 328 iterations at $\alpha = 0.1$ for marginally better scores. This makes $\alpha = 0.5$ a practical default, while smaller values may be preferred when utility is prioritized over computational cost.

4.6 Impact of image resolution

To evaluate EDIALS' scalability, we investigated the effect of reducing image resolution on utility preservation and computational efficiency. We re-trained our AAE at 128×128 and 64×64 resolutions and compared the results with the standard implementation employed so far, which uses 256×256 images.

Reducing resolution offers significant computational benefits with varying impacts on utility preservation. Downsampling to 128×128 enabled batch processing of 8 images simultaneously, reducing processing time by 73% (from 33.0s to 9.0s per image). Further reduction to 64×64 allowed for batches of 16 images, decreasing processing time by 90% (to 3.2s per image). These improvements stem from the quadratic reduction in memory requirements, which enables more efficient parallel processing.

Table 6: Resolution impact on EDIALS runtime and utility at the lower bound RIR (0.42%). Reduced resolutions enable batch processing, decreasing runtime by 73% (128×128) and 90% (64×64) relative to 256×256, with consequent reductions of demographic classification accuracy across tasks.

Resolution	Batch size	Runtime (s)	F_1 Score (%) Age	Gender	Race
256×256	1	33.0	84.66	97.91	82.61
128×128	8	9.0	81.24	94.53	78.42
64×64	16	3.2	72.38	88.65	70.14

The impact on utility varies by classification task. Gender classification demonstrated remarkable resilience to resolution reduction, with only a minor degradation at 128×128 (3.38%) and moderate impact at 64×64 (9.26%). Age and race classifications showed greater sensitivity, with modest degradation at 128×128 (3.42% and 4.19%, respectively) but more substantial declines at 64×64 (12.28% and 12.47%). As observed in Section 4.2, gender classification generally maintains better performance under image modifications because it relies on broader facial structure.

The decline in F_1 scores at lower resolutions reflects the compound effect of information loss in our pipeline. While high-frequency details persist at higher resolutions, they become increasingly compromised at lower resolutions [41]. When we apply targeted distortion to these already resolution-compromised features, the effect particularly impacts tasks requiring fine detail discrimination, especially at 64×64 resolution, where age and race classification performance deteriorates significantly. These findings indicate that 128×128 resolution offers a good compromise between efficiency and utility preservation. When time constraints are not critical, the 256×256 resolution offers better utility preservation.

Importantly, when considering inference time alone, the 128×128 resolution version achieves processing times comparable to DeepPrivacy2 while maintaining better RIR values and utility metrics.

5 Conclusion, limitations and future work

We presented EDIALS, a novel approach to facial image anonymization that operates in the latent space using explainability techniques to selectively target identity-revealing features. Our empirical evaluation confirms that EDIALS achieves a superior balance between privacy and utility compared to conventional methods and GAN-based approaches, achieving anonymization at the theoretical minimum of random guessing (RIR = 0.42%) while maintaining high utility for demographic classification tasks (scores F_1: 84.66% for age, 97.91% for gender and 82.61% for race).

A limitation of EDIALS is its higher computational complexity compared to DeepPrivacy2, as it requires both the AAE and black-box models to remain on the GPU, precluding efficient batch processing at higher resolutions. When utility preservation is critical, this performance cost is justified, and can be reduced by processing

at lower resolutions, which achieves a better privacy-utility balance than DeepPrivacy2.

Future research could improve EDIALS by: (1) exploring adversarial training techniques to further strengthen privacy against reconstruction attacks; (2) investigating disentangled representations that better separate identity from demographic attributes; and (3) exploring theoretical privacy guarantees for latent space modifications that could complement the empirical privacy.

Acknowledgments

This research was funded by the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and D. Sánchez), MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe" under grant PID2024-157271NB-I00 (CLEARING-IT) and the European Commission (project HORIZON-101292277 "SoBigDataIP"). It was also partially supported by the European Union's Horizon 2020 programme under grant agreement No. 101120763 (TANGO) and grant agreement No. 834756 (XAI). Y. Khan acknowledges financial support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 945413 and from the Universitat Rovira i Virgili. This paper reflects only the authors' views, and neither the European Union nor the granting authorities can be held responsible for any use that may be made of the information it contains.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 965–974.

[3] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. 2022. A critical review on the use (and misuse) of differential privacy in machine learning. *Comput. Surveys* 55, 8 (2022), 1–16.

[4] George EP Box and Mervin E Muller. 1958. A note on the generation of random normal deviates. *The annals of mathematical statistics* 29, 2 (1958), 610–611.

[5] Gary Bradski. 2000. The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25, 11 (2000), 120–123.

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] Juan J Cerrolaza, Antonio R Porras, Awais Mansoor, Qian Zhao, Marshall Summar, and Marius George Linguraru. 2016. Identification of dysmorphic syndromes using landmark-specific local texture descriptors. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1080–1083.

[8] Jia-Wei Chen, Li-Ju Chen, Chia-Mu Yu, and Chun-Shien Lu. 2021. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6478–6487.

[9] Tim Daniels. 2024. Photography Statistics & Trends (2024). https://www.lapseoftheshutter.com/photography-statistics/?utm_source=chatgpt.com.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[13] Anil Egin, Andrea Tangherloni, and Antitza Dantcheva. 2025. Now You See Me, Now You Don't: A Unified Framework for Expression Consistent Anonymization in Talking Head Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5925–5934.

[14] Liyue Fan. 2018. Image pixelization with differential privacy. In *Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo, Italy, July 16–18, 2018, Proceedings 32*. Springer, 148–162.

[15] Liyue Fan. 2019. Differential privacy for image publication. In *Theory and Practice of Differential Privacy (TPDP) Workshop*, Vol. 1. 6.

[16] Ian H Fraser, Greg L Craig, and Denis M Parker. 1990. Reaction time measures of feature saliency in schematic faces. *Perception* 19, 5 (1990), 661–673.

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

[18] Riccardo Guidotti, Anna Monreale, Stan Matwin, and Dino Pedreschi. 2020. Black box explanation by learning image exemplars in the latent feature space. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, 189–205.

[19] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems* 30 (2017).

[20] Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M Krawitz, Susanne B Kamphausen, Martin Zenker, et al. 2019. Identifying facial phenotypes of genetic disorders using deep learning. *Nature medicine* 25, 1 (2019), 60–64.

[21] Rami Haffar, David Sánchez, and Josep Domingo-Ferrer. 2025. Multi-Task Faces (MTF) Data Set: A Legally and Ethically Compliant Collection of Face Images for Various Classification Tasks. *IEEE Access* (2025).

[22] Rami Haffar, David Sánchez, Younas Khan, and Josep Domingo-Ferrer. 2025. Explainability-Driven Incremental Image Anonymization. *Transactions On Data Privacy* (2025).

[23] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

[24] Håkon Hukkelås and Frank Lindseth. 2023. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1329–1338.

[25] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*. Springer, 565–578.

[26] Seyyed Mohammad Sadegh Moosavi Khorzooghi and Shirin Nilizadeh. 2023. Examining StyleGAN as a utility-preserving face de-identification method. *Proceedings on Privacy Enhancing Technologies* (2023).

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[28] Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. 2025. Face anonymization made simple. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1040–1050.

[29] Erich Leo Lehmann and George Casella. 1998. *Theory of point estimation*. Springer.

[30] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 34–42.

[31] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* (2017).

[32] Bryson Lingenfelter, Sara R Davis, and Emily M Hand. 2022. A quantitative analysis of labeling issues in the celeba dataset. In *International Symposium on Visual Computing*. Springer, 129–141.

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.

[34] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11976–11986.

[35] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5447–5456.

[36] Ibomoiye Domor Mienye and Theo G Swart. 2025. Deep autoencoder neural networks: A comprehensive review and new perspectives. *Archives of computational methods in engineering* (2025), 1–20.

[37] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[38] Carman Neustaedter, Saul Greenberg, and Michael Boyle. 2006. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 1 (2006), 1–36.

[39] Elaine M Newton, Latanya Sweeney, and Bradley Malin. 2005. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering* 17, 2 (2005), 232–243.

[40] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International conference on machine learning*. Pmlr, 1310–1318.

[41] Ashay Patel, Michela Antonelli, Sebastien Ourselin, and M Jorge Cardoso. 2025. Resolution Invariant Autoencoder. *arXiv preprint arXiv:2503.09828* (2025).

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
[43] Lutz Prechelt. 2002. Early stopping-but when? In *Neural Networks: Tricks of the trade*. Springer, 55–69.
[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[45] Erik A Roberts, Chelsea Troiano, and Jeffrey H Spiegel. 2016. Standardization of guidelines for patient photograph deidentification. *Annals of plastic surgery* 76, 6 (2016), 611–614.
[46] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*. 10–15.
[47] Javid Sadr, Izzat Jarudi, and Pawan Sinha. 2003. The role of eyebrows in face recognition. *Perception* 32, 3 (2003), 285–293.
[48] Modesto Castrillón Santana, Oscar Déniz-Suárez, Luis Antón-Canalís, and Javier Lorenzo-Navarro. 2008. Face and Facial Feature Detection Evaluation - Performance Evaluation of Public Domain Haar Detectors for Face and Facial Feature Detection. In *VISAPP 2008: Proceedings of the Third International Conference on Computer Vision Theory and Applications, Funchal, Madeira, Portugal, January 22-25, 2008 - Volume 2*, Alpesh Ranchordas and Helder Araújo (Eds.). INSTICC - Institute for Systems and Technologies of Information, Control and Communication, Setúbal, Portugal, 167–172.
[49] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. 2015. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence* 38, 7 (2015), 1439–1451.
[50] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc. IEEE* 94, 11 (2006), 1948–1962.
[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.
[52] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 9677–9696.
[53] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* 83 (2022), 19–52.
[54] Qiang Alex Zhao and John T Stasko. 1998. Evaluating image filtering based techniques in media space applications. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*. 11–18.

A Implementation Details for Reproducibility

To facilitate reproducibility, this appendix provides comprehensive implementation details for the EDIALS framework, including precise architectural specifications, training hyperparameters, and algorithmic procedures that complement the open-source codebase available in our GitHub repository: https://github.com/yk0803/EDIALS.

A.1 Adversarial Autoencoder (AAE) Architecture

Encoder ($\mathcal{E}$): The encoder is based on a truncated ConvNeXt-Base model [34] pre-trained on ImageNet [10] with classification layers removed. It outputs a 1024-dimensional bottleneck representation to compress image information while preserving semantic details.

Decoder ($\mathcal{D}$): The decoder implements a progressive upsampling architecture with five transposed convolutional stages that gradually increase spatial resolution from 8×8 to 256×256 pixels. Each upsampling stage incorporates transposed convolution followed by LayerNorm [1] and GELU activation [23]. The final layer employs a tanh activation to generate normalized RGB images within the $[-1, 1]$ range.

Discriminator (δ): The discriminator employs a convolutional network with five convolutional layers that process latent vectors, followed by fully connected layers. It uses LayerNorm, GELU activations, and dropout [51] for regularization. The discriminator outputs a sigmoid probability for latent vector validity with threshold $\delta(h) \geq 0.5$ and uses hinge loss [31] with gradient penalty regularization [19] for stable adversarial training.

A.2 Decision Tree and Explainability Framework

EDIALS uses Scikit-learn DecisionTreeClassifier [42] with Gini impurity criterion [6], maximum depth of 128, and pruning mechanisms to handle contradictory rules. It is trained on neighborhood H and corresponding labels Y to approximate black-box model decisions in latent space. Features $Z_r = \{z_1, z_2, \ldots, z_m\}$ are extracted from root-to-leaf paths where splitting conditions identify latent dimensions that significantly influence identity classification.

EDIALS integrates ABELE framework [18] for model-agnostic explanation of black-box decisions in latent feature space. It leverages ABELE's methodology for encoding images into latent space and extracting decision rules, focusing on factual rules rather than counterfactual rules for controlled image perturbation.

A.3 Neighborhood Generation

The implementation generates 200-1000 latent vectors per input using Gaussian noise sampling [4]. The *neighgen* module creates instances with both matching ($H_=$) and non-matching ($H_\neq$) classifications: $H = H_= \cup H_\neq$. This balanced neighborhood enables the decision tree to learn the local behavior of the black-box model accurately. Generated latent vectors are filtered using the discriminator δ to retain only valid vectors ($\check{H}$) through the PyTorchImageRandomAdversarialGeneratorLatent class.

A.4 Training and Distortion Parameters

Training Configuration: We train the AAE for 200 epochs with early stopping [43] after 5 consecutive epochs without validation improvement. We use Adam optimizer [27] with learning rate 0.0001 for encoder/decoder and 0.00005 for the discriminator. Mean Squared Error (MSE) reconstruction loss [29] weight is 10.0, adversarial loss weight is 1.0. Gradient clipping [40] is applied with maximum norm 0.5.

Distortion Process: It employs iterative feature modification with distortion factor $\alpha = 0.5$, brightness normalization, and latent norm scaling. The direction of distortion is determined by these extracted conditions: if a rule condition specifies $z_i <$ threshold, then α is added to increase the feature value and violate this condition; conversely, if a rule condition states $z_i \geq$ threshold, then α is subtracted to push the feature value below the threshold. Features are processed iteratively until black-box model misclassifies ($b(\mathcal{D}(z_{\text{distorted}})) \neq b(x)$), which indicates successful anonymization.