PerceptAnon: Exploring the Human Perception of Image Anonymization Beyond Pseudonymization for GDPR

Kartik Patwari [*†1] **Chen-Nee Chuah** [1] **Lingjuan Lyu** [2] **Vivek Sharma** [*⋆2]

Abstract

Current image anonymization techniques, largely focus on localized pseudonymization, typically modify identifiable features like faces or full bodies and evaluate anonymity through metrics such as detection and re-identification rates. However, this approach often overlooks information present in the entire image post-anonymization that can compromise privacy, such as specific locations, objects/items, or unique attributes. Acknowledging the pivotal role of human judgment in anonymity, our study conducts a thorough analysis of perceptual anonymization, exploring its spectral nature and its critical implications for image privacy assessment, particularly in light of regulations such as the General Data Protection Regulation (GDPR). To facilitate this, we curated a dataset specifically tailored for assessing anonymized images. We introduce a learning-based metric, **PerceptAnon**, which is tuned to align with the human **Percept**ion of **Anon**ymity. PerceptAnon evaluates both original-anonymized image pairs and solely anonymized images. Trained using human annotations, our metric encompasses both anonymized subjects and their contextual backgrounds, thus providing a comprehensive evaluation of privacy vulnerabilities. We envision this work as a milestone for understanding and assessing image anonymization, and establishing a foundation for future research. The codes and dataset are available in https://github.com/SonyResearch/gdpr_perceptanon.

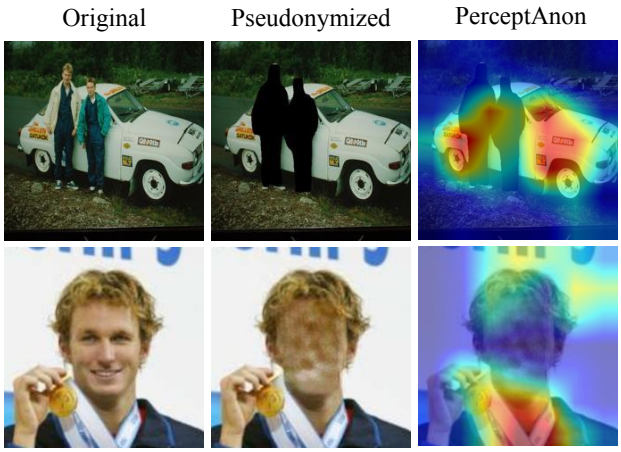


Figure 1. Pseudonymization focuses on anonymizing local regions that can leave privacy-compromising cues in the image. PerceptAnon focuses on identifying and quantifying such cues.

1. Introduction

The vast proliferation of data and digital platforms has heightened data privacy concerns, necessitating effective anonymization strategies to protect Personally Identifiable Information (PII) in images. The General Data Protection Regulation (GDPR) distinguishes 'pseudonymisation'– modifying personal data to prevent direct attribution to individuals without additional information—with 'anonymization', where personal identification becomes unfeasible (EDPS, 2021). Under GDPR, anonymity is achieved when data is processed such that the subject becomes unidentifiable, with masking cited as the most feasible option for images (Weitzenboeck et al., 2022; Barta, 2018). Current image anonymization techniques, such as masking, blurring, pixelation (Du et al., 2019; Yang et al., 2022), and generation (Li et al., 2021), largely align with pseudonymization, focusing on modifying identifiable attributes like faces or full bodies.

Traditional research in this domain has mostly concentrated on localized areas within images, using detection and re-identification rates for anonymity assessment. GDPR guidelines mention that "*the use of additional information can lead to the identification of individuals*" (EDPS, 2021) like recognizable items or locations. As shown in Figure 1, addi-

[*]Equal contribution [†]Work done while interning at Sony AI [⋆]VS started and led the project [1]Department of Electrical and Computer Engineering, University of California Davis, CA, USA [2]Sony AI. Correspondence to: Vivek Sharma <viveksharma@sony.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

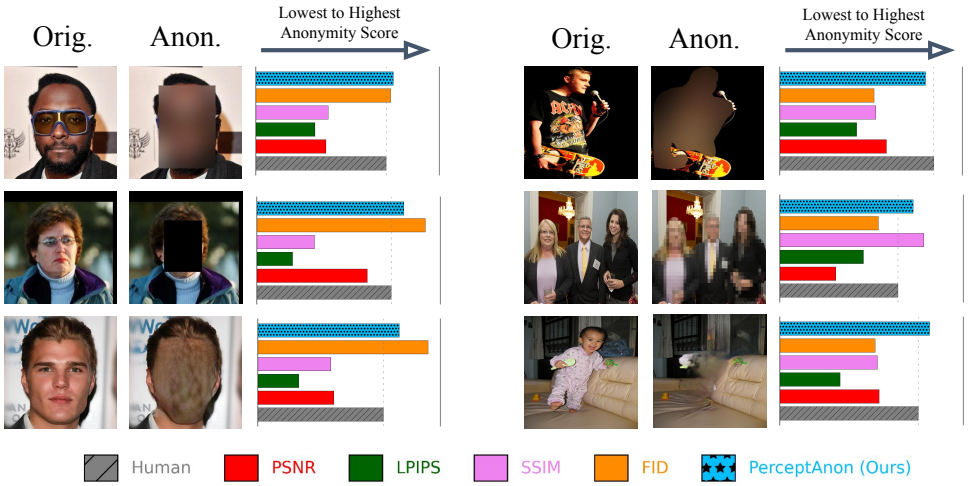


Figure 2. Comparison of traditional metrics and PerceptAnon against human anonymity assessments in image-anonymized pairs. The graph ranks anonymity scores from lowest to highest. PerceptAnon shows the closest alignment with human judgments. All scores were normalized between 0 and 1 for visual comparison, with PSNR, SSIM, and FID inverted for consistency with anonymity criteria.[2] Notably, human and PerceptAnon scores are based off only the anonymized image (without original reference). Best viewed in color.

tional recognizable and privacy-compromising cues exist in pseudonimized image, as local detectability tends to overlook the full image content. We advocate that considering potential privacy-leaking cues throughout the entire image is crucial for holistic assessment, crucial in real-world contexts where privacy concerns can extend beyond the direct detectability. Moreover, detectability evaluations are often binary—either re-identified or not—ignoring the spectrum of effectiveness that characterizes perceptual anonymization.

In response, our study focuses on the post-detection phase of anonymization techniques, acknowledging that the entire image, including the background and ancillary details, can compromise privacy. Our research is inspired by recent work in adopting a human-centric metric for privacy evaluation of reconstructed images (Sun et al., 2023). We utilize standard image assessment metrics like Structural Similarity Index Measure (SSIM) or Learned Perceptual Image Patch Similarity (LPIPS), which have been recently evaluated in human perception-based image comparison and privacy assessment contexts (Fu et al., 2023; Sun et al., 2023) and evaluate their alignment with human perception of anonymity. In Figure 2, we compare various original-anonymized image pairs, analyzing how well traditional image assessment metrics align with human assessments of anonymity. We find a lack of strong alignment between traditional metrics and human perception, primarily due to their deficiency in capturing semantic aspects of anonymity. This observation highlights the need for a new metric that more closely reflects human judgment.

To achieve this, we curate a new dataset which consists of images from widely-used vision datasets in the detection and anonymization domain, specifically for persons (MS-COCO (Lin et al., 2014), PASCAL VOC (Everingham et al., 2010)) and faces (LFW (Huang et al., 2008), CelebA-HQ (Liu et al., 2015)). We then anonymize selected images using various existing methods. We propose two annotation approaches: one where evaluators assess only the anonymized images, and another where they compare the anonymized images with their originals. This allows us to explore the ability of humans to gauge anonymity from a standalone anonymized image, how the presence of the original image influences this perception, and a comprehensive understanding of privacy vulnerabilities. Using mean human annotations we compare the effectiveness of existing image comparison metrics against human judgments, highlighting their degree of alignment or disparity.

Our work introduces *PerceptAnon*, a novel metric that aligns better with human perceptions. Our metric is not only trained to evaluate original-anonymized image pairs but also focuses on solely anonymized images. PerceptAnon represents a shift towards a holistic evaluation of anonymization, assessing the cumulative image rather than just the anonymized subjects. We conduct experiments to quantify whether learning anonymity from our annotations is more effective as a classification or regression task, and explore the optimal granularity for aligning with human perception.

The key contributions of this paper are as follows:

- To the best of our knowledge, this is the first work that attempts to quantify image anonymization specifically from a human-centric perspective.
- We curate a novel dataset, combining existing datasets with images anonymized by prevalent methods, scored

[2] PSNR, SSIM, and FID scores are inverted because higher values denote more similarity, thus less anonymity, contrary to our anonymity assessment criteria.

for anonymity by annotators.
- We introduce PerceptAnon, a learned metric mirroring human perception for assessing anonymized images.
- We investigate the optimal approach to understanding and learning image anonymization, whether it should be treated as a regression or classification problem and what granularity.

2. Related Works

Anonymization Techniques. Early image anonymization relied on traditional obfuscation methods like masking, blurring, and pixelation (Du et al., 2019; Yang et al., 2022), which are prevalent in real-world applications (Pachni A., 2022; Jaichuen et al., 2023). Despite their widespread use, these techniques have notable limitations, especially in the context of advanced de-obfuscation methods (Zhang et al., 2020; Vishwamitra et al., 2017; Oh et al., 2016; McPherson et al., 2016). The rise of deep learning introduced generative models, for targeted face and full-body anonymization (He et al., 2023; Rosenberg et al., 2023; Hukkelås et al., 2019). This includes attribute-preserving approaches (Li et al., 2021; Barattin et al., 2023; Hellmann et al., 2023) and recent advancements in photo-realistic full-body anonymization (Maximov et al., 2020; Hukkelås et al., 2023). Inpainting techniques have also can also be used to "erase" sensitive areas effectively (Upenik et al., 2019; Cao et al., 2021). However, the efficacy of all these methods is predominantly assessed on anonymized subjects, often overlooking the comprehensive visual context within images which is critical for privacy. Our work addresses this gap by introducing a metric to evaluate anonymity leakage in a holistic manner, factoring in the entire image rather than focusing solely on the anonymized subjects.

Contextual Cues in Anonymity Assessment. Recent works (Nagrani et al., 2018; Qian et al., 2017; Shao et al., 2019) indicate the potential of auxiliary information, like voice or social network data, in compromising anonymity. However, the specific role of visual elements within images remains less explored. The work by (Hukkelås & Lindseth, 2023b) focuses on how anonymization affects the recognition of other objects in images during training and testing, but does not consider the role of surrounding information in quantifying the effectiveness of anonymization. In contrast, our work investigates this aspect.

Perceptual Metrics. DreamSim (Fu et al., 2023) is a metric designed to assess image similarity. Given a set of three images, it determines which one, out of two options, is more similar to the third reference image. SemSim (Sun et al., 2023) proposes a learned metric that integrates human judgment for evaluating privacy in the context of image reconstruction attacks. Our research, inspired by these works, shifts focus to a different aspect of privacy: the quantification of anonymization.

3. Quantifying Image Anonymization

In this section, we explore current image anonymization approaches, discuss the limitations of pseudonymization techniques and evaluation metrics. Then, we investigate how to capture anonymity throughout entire images. Finally, we discuss the potential of human perception in aiding the evaluation of image anonymization.

3.1. Current Anonymization Landscape

Current image anonymization techniques (Barattin et al., 2023; Hukkelås & Lindseth, 2023a), primarily employ pseudonymization methods like blurring or generating regions to obscure individual PII. While effective in localized contexts, traditional metrics such as re-identification (re-ID) or detection rates do not fully capture privacy implications in the broader image context, including background details and interactions. This highlights the need for a more holistic metric that encompasses the entire image content.

3.2. Problem Statement

The inherent challenge in image anonymization is not just preserving privacy by obscuring individual PII but ensuring no observable privacy leaks are present across the entire image. Our research aims to address these questions:

1. How can we assess total anonymity of an image holistically – beyond just pseudonymization?
2. How to learn human perception of anonymity, viewing it as regression or (ordinal) classification?
3. What is an appropriate scale for measuring anonymity: binary, or continuous and at what granularity?
4. Should anonymized images be assessed independently or in comparison with their original counterparts?

3.3. Pitfalls of Existing Metrics

The metrics used in current anonymization techniques (Barattin et al., 2023; Hukkelås & Lindseth, 2023a; Maximov et al., 2020) , such as Re-ID or detection rates, primarily assess the effectiveness of pseudonymization in erasing specific PII like faces or bodies. While effective for these specific tasks, these metrics fall short in broader contexts. Firstly, they are binary (detected/not detected, re-identified/not re-identified), oversimplifying the complexity of anonymity. Secondly, these metrics are designed for localized feature assessment and overlook the wider image context rich in de-anonymizing cues. Crucial elements like location-specific backgrounds, unique scene objects, or subject interactions within the image are often neglected. Therefore, while Re-ID and detection rates serve well for pseudonymization assessment of specific PII, they are not suitable for holistic evaluation of image anonymity, emphasizing the need for more comprehensive metrics.

3.4. Evaluating Entire Images

To the best of our knowledge, this is the first work focusing on the anonymity evaluation of the entire image beyond pseudonymization, and currently, there exist no metrics specifically designed for this task. To form a baseline, we draw from adjacent realms such as the privacy assessment of reconstructed images (Sun et al., 2023) which commonly utilize traditional image quality and similarity metrics. Metrics like Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), and Fréchet Inception Distance (FID), while not initially developed for assessing anonymity, offer an objective measure of visual changes between original and anonymized image pairs. However, their limitation is apparent in their inability to identify which changes are critical for preserving anonymity. Our proposed metric, PerceptAnon, is designed to overcome this limitation by providing a holistic, perception-based assessment.

3.5. Human Perception for Anonymization

Recognizing the complexity of total image anonymity, we rely on human perception as a core principle. Human observers, capable of assessing anonymity without a reference image, consider a combination of features and contexts. This holistic approach is crucial for scenarios where anonymity extends beyond obscuring PII to encompassing all identifiable traits. Our metric, designed to mirror this nuanced human perception, complements existing evaluations like Re-ID and detection rates, adding a layer that accounts for broader and more subjective aspects of visual privacy. We recognize the complexity and multifaceted nature of human perception in the context of anonymity, acknowledging that it is often difficult to quantify (Wiles et al., 2012; Saunders et al., 2015). Our work presents an initial step towards addressing this challenge.

4. PerceptAnon: A Human-Centric Metric

Recognizing human perception's key role, this section details our methodology for quantifying image anonymization, starting with curating a dataset and annotations, focused on new scopes and setups. We then detail the training of PerceptAnon, a metric trained on these annotations.

4.1. Curating a Dataset for Anonymization

Typically anonymization consists of two stages. The first is the PII detection/segmentation and second is applying an anonymization technique; our work focuses on analyzing the post detection phase. Existing datasets lack human-labeled data that focus on the holistic assessment of image anonymization. To address this, we introduce a unique dataset that encompasses a wide spectrum of anonymization, from pseudonymization (local) to full anonymization (global). This dataset aims to evaluate human perception of anonymization in various contexts, particularly focusing on faces and full bodies – areas primarily researched in previous studies (Barattin et al., 2023; Yuan et al., 2022; Yang et al., 2022; Hukkelås & Lindseth, 2023a).

Our dataset includes 500 images each from COCO, VOC, LFW, and CelebA-HQ, chosen for their relevance in face and full-body anonymization. COCO and VOC provide people in a diverse range of backgrounds, crucial for evaluating (global) anonymization, while CelebA and LFW are pivotal in face anonymization research (Barattin et al., 2023; Peng et al., 2022; Yuan et al., 2022). Each image underwent four anonymization techniques: masking, blurring/pixelation, inpainting, and generation. To illustrate the spectrum of anonymization, we generated anonymized images with two scopes: *Local* (akin to pseudonymization) where anonymized image only focus on PII regions (face, persons) and *Global* where anonymized images retain background content with PII sanitized. Sample anonymized images can be found in Appendix A. To the best of our knowledge, no prior work has comprehensively investigated all these techniques on both faces and full bodies across such a diverse set of datasets, including realistic contexts like COCO and VOC, going beyond mere pseudonymity.

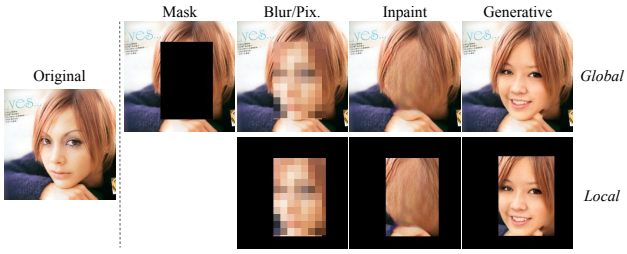


Figure 3. Local vs. Global Scopes. Local views focus solely on anonymized PII, excluding background. Global only sanitizes PII but retains background. Local mask is not considered.

Anonymization technique details. DeepFillV2 (Yu et al., 2019) and DeepPrivacy2 (Hukkelås & Lindseth, 2023a) are used for inpainting and generating respectively. Blurring and pixelation, are treated as similar obfuscation effects; hence, we pool them and randomly blurred half of the original images and pixelated the remainder. Pixelation was executed with a block size of 16×16, in line with (Hukkelås et al., 2019). For blurring, we applied a Gaussian blur filter with a radius equal to 1/8 of the width of the bounding box (Dietlmeier et al., 2021). Thus, we generated four anonymized versions for each original image.

We utilized DSFD (Li et al., 2019) with ResNet-152 backbone for face detection, which achieved AP scores of 96.6, 95.7, and 90.4 on the WIDER Face (Yang et al., 2016) easy, medium, and hard sets, respectively. Full-body detection

and segmentation were performed using Mask-RCNN (He et al., 2017), which has a box mAP of 47.4 and mask mAP of 41.8 on the COCO 2017 validation set. Both detectors are popular choices in previous anonymization research (Hukkelås et al., 2019; Hukkelås & Lindseth, 2023a).

Scope details. To differentiate and analyze both pseudonymization and anonymization, we present two image variants for each anonymization technique: (1) *Local*, which concentrates solely on PII, aligning with pseudonymization, and (2) *Global*, which retains the original context but alters PII, embodying anonymization. Figure 3 illustrates both global and local image examples. Additionally, Figure 4 shows the distribution of human annotator scores for each method and scope. Notably, global images, which include background context, consistently receive lower anonymity scores than local images. This trend emphasizes the significance of context and background in the overall perception of anonymity.

4.2. Human Annotations

Annotation setups. Our dataset introduces two annotation setups; *Human Annotation 1* ($HA1$): Annotators view only the anonymized image, excluding generated images to prevent misjudgments without original image comparison. *Human Annotation 2* ($HA2$): Annotators see the original and its anonymized counterpart, excluding global masked images to avoid obvious recognition. Local mask images are excluded in both setups due to complete image obscuration. Shared anonymized images between the two setups were those processed with blur/pixelate and inpaint techniques.

Our dataset 500 original images, each is anonymized using four different techniques and two scopes, resulting in eight variations per image and thus 4000 images total. For $HA1$ setup, we accumulated a total of 2500 annotations across five categories: Global-Mask, Local-Blur, Global-Blur, Local-Inpaint, and Global-Inpaint. In $HA2$ setup, we gathered 3000 annotations spanning six categories: Local-Blur, Global-Blur, Local-Inpaint, Global-Inpaint, Local-Generative, and Global-Generative. Our dataset scale is consistent with similar human perceptual metric studies (Sun et al., 2023). Our dataset, uniquely the first for global image privacy assessment with human annotations, covers all four major anonymization methods for faces and bodies, providing a comprehensive scope. We utilize various train/test strategies (as descibed in Section 5.1), to further combat the dataset scale and study robustness and generalizability.

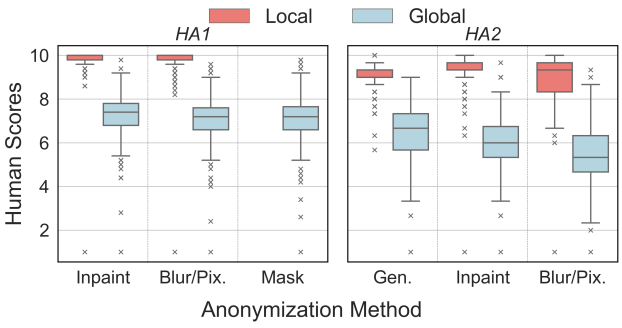


Figure 4. Distribution of human scores by anonymization method and scope (global/local). Generative method was not considered in $HA1$, and Mask was not in $HA2$.

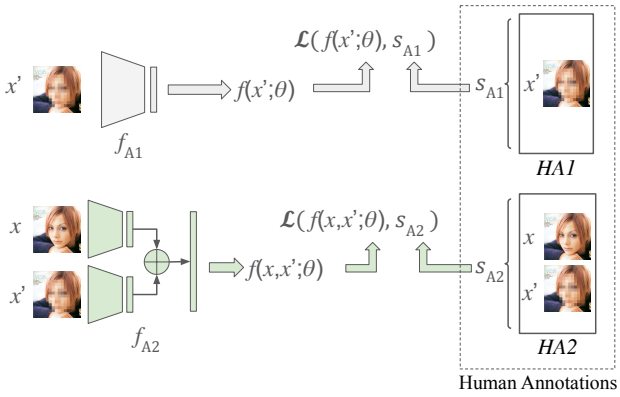


Figure 5. PerceptAnon Training Pipelines: $HA1$ utilizes a CNN (f_{A1}) processing anonymized images (x') with $HA1$ scores (s_{A1}); $HA2$ employs a Siamese network (f_{A2}) for original-anonymized pairs (x, x') with $HA2$ scores (s_{A2}). Loss function $\mathcal{L}$ is MSE or CE for regression or classification, respectively.

Collecting human annotations. Following a similar setup of (Sun et al., 2023), our images were annotated by 6 independent annotators. Each annotator was asked to score the anonymity achieved on a scale of 1-10, with 10 being the highest degree of anonymity. Final scores were averaged across all annotators for each image or image pair. This approach lends itself to regression analysis using mean human scores, with further details on thresholding and granularity testing described in Section 5.1. We measured annotator consistency using Cronbach's Alpha (CA), where a higher score denotes greater annotation reliability. Our annotations yielded a CA value of 0.87, which is comparable to vision datasets that involve subjective, score-based annotations (Song et al., 2015; Gygli et al., 2014).

4.3. Computing PerceptAnon Metric

Inspired by recent learning-based perception metrics (Sun et al., 2023; Fu et al., 2023), we introduce PerceptAnon, a novel metric designed to capture human perceptions of image anonymization. PerceptAnon leverages human-annotated data, aiming to reflect a more intuitive and human-centric understanding of anonymization. PerceptAnon utilizes Convolutional Neural Networks (CNNs) to learn from human annotations. CNNs not only provided a robust basis for our studies but also facilitate a straightforward and in-

terpretable approach that aligns with our study exploration. We provide further discussion on our choice of using CNNs in Appendix B.

Training. Figure 5 demonstrates our training pipelines. Given an original image x and its anonymized counterpart x', the model is trained on annotation scores s, where s is between 0 and 1, derived by normalizing the raw human ratings on a scale of 1-10. The supervised regression model $f(x', s; \theta)$ predicts an anonymization score that aligns with human evaluations. To explore both regression and classification, we transform the continuous scores s into ordinal categories, using rounding or thresholding for discretization (details in Section 5.1).

For $HA1$, the model $f_{A1}(x', s_{A1}; \theta)$ is trained using s_{A1} scores, replicating scenarios where only anonymized images are available, without reference originals. This setup provides insights into anonymization perception in isolation. For $HA2$, we use a Siamese network architecture, $f_{A2}(x, x', s_{A2}; \theta)$, where we concatenate features of both the original and anonymized images, and train end-to-end on s_{A2} scores.

Observations and limitations. PerceptAnon aims to encompass a broad spectrum of anonymization, from localized pseudonymization to comprehensive anonymization. Despite its simplicity, it is found to be effective under varying scenarios (see Section 5). Key to its effectiveness is its training on human annotations, enabling it to discern nuances in anonymization beyond what traditional pixel-level or patch-based metrics capture. However, a limitation lies in its reliance on annotated data quality and diversity.

5. Experiments

In this section, we first introduce the experiment setup and implementation details of our proposed approach. We then assess existing metrics alignment with human perception and demonstrate the effectiveness of PerceptAnon. Finally we perform additional ablation studies and discussions to evaluate image anonymity with PerceptAnon.

5.1. Experiment Setup

Training/Testing setups. We evaluate with three distinct training and testing strategies utilizing our proposed anonymization dataset: (1) *Whole Dataset (All)* – the full dataset, split 60:20:20 for training, validation, and testing; (2) *Leave-One-Out-Validation (LOOV)* – each of the four source datasets (COCO, VOC, LFW, CelebA) is used in turn as a test set, with the others for training; and (3) *Task-Specific* – separate person (*Task-Person*) and face (*Task-Face*) focused evaluations, with respective images from COCO and VOC, and LFW and CelebA, split in a 60:20:20 ratio. This approach aims to assess the effectiveness of our methods in both face and full-body anonymization scenarios, and additionally, to test the model's ability to generalize to unseen datasets.

Granularity of anonymity. Our annotation scores ranging from 1 to 10 (with 10 being the highest anonymity), were normalized between 0 and 1 for regression analysis, termed Regression (Mean). For classification, we employed various thresholding strategies: 10-class (rounding mean scores), 5-class (binning every two scores), 3-class (grouping scores at equal intervals into low, medium, high), and binary (dividing into low and high anonymity levels). Each classification model's thresholded scores were also normalized, suitable for regression.

Backbone architectures. We experiment with the following backbone architectures: ResNet18,50,152 (He et al., 2016), DenseNet121 (Huang et al., 2017), and AlexNet (Krizhevsky et al., 2012). For all architectures, we initialize with ImageNet pretrained weights. We use ResNet50 for main evaluation (see Appendix B for results on other architectures).

Evaluating metrics. We use Spearman's (ρ) and Kendall's (τ) rank correlations between predicted and ground truth test scores across various label types. These non-parametric rank correlations, suitable for our ordinal data, range from -1 to 1, with values near the extremes indicating strong model consistency with human annotations. For models trained on transformed labels (e.g., binary), evaluation is similarly conducted against correspondingly transformed test scores. Following SemSim (Sun et al., 2023), we also evaluate the correlation of human scores with traditional image assessment metrics. Since we use rank-based correlations, the values do not need to be in the same distribution and range. We do not compare against SemSim in our evaluations, as our focus is on anonymization, whereas SemSim is designed for assessing in image reconstruction contexts.

Comparison with traditional metrics. For a comprehensive evaluation, we incorporated traditional image assessment metrics alongside our human-centric PerceptAnon metric. Specifically, we utilized implementations from the DeepPrivacy (Hukkelås et al., 2019) for calculating MSE, PSNR, LPIPS, SSIM, and FID. Notably, the FID metric uses the InceptionV3 (Szegedy et al., 2016) architecture, while the LPIPS metric was based on the AlexNet (Krizhevsky et al., 2012) architecture. Further details and metric descriptions can be found in Appendix F.

5.2. Implementation Details

We utilized PyTorch (Paszke et al., 2019) framework and its default ImageNet pretrained models for our implementations. Models were trained on NVIDIA RTX A4500 GPUs, each with 20GB of memory. In main evaluation we use a

Train/Test Setup	Metrics	PSNR	MSE	LPIPS	SSIM	FID	PerceptAnon (Ours)
All	ρ	-0.7011	0.7011	0.7675	-0.8358	0.6578	**0.8817**
	τ	-0.5018	0.5018	0.5544	**-0.7601**	0.4667	0.7119
LOOV-VOC	ρ	-0.7448	0.7448	0.8244	-0.8185	0.6995	**0.8603**
	τ	-0.5437	0.5437	0.6288	-0.6289	0.5095	**0.6570**
LOOV-COCO	ρ	-0.771	0.771	0.805	-0.7702	0.733	**0.8643**
	τ	-0.5649	0.5649	0.6051	-0.5712	0.5385	**0.6845**
LOOV-LFW	ρ	-0.7354	0.7354	0.7574	-0.7615	0.7289	**0.8278**
	τ	-0.5256	0.5256	0.5487	-0.5509	0.5141	**0.6353**
LOOV-CelebA	ρ	-0.6239	0.6239	0.7301	-0.7321	0.6634	**0.8478**
	τ	-0.4407	0.4407	0.5151	-0.518	0.4594	**0.6549**
Task-Person	ρ	-0.7313	0.7313	0.7909	-0.75	0.6858	**0.8831**
	τ	-0.524	0.524	0.5929	-0.5452	0.4971	**0.7120**
Task-Face	ρ	-0.7547	0.7547	0.7906	-0.7838	0.7447	**0.8774**
	τ	-0.5528	0.5528	0.5887	-0.5825	0.547	**0.6940**

Table 1. Correlation of various metrics with human anonymity mean score – Regression (mean) on $HA1$, assessed using Spearman's (ρ) and Kendall's (τ) correlations. PerceptAnon is trained via regression on mean human scores. Results reflect test set correlations across three strategies: the entire dataset ('All'), Leave-One-Out Validation (LOOV), and task-specific setups (person or face). LOOV-VOC indicates VOC as the test dataset and remaining as train, and Task-Person for person-anonymized images only.

Train/Test Setup	Metrics	PSNR	MSE	LPIPS	SSIM	FID	PerceptAnon (Ours)
All	ρ	-0.7631	0.7631	0.7622	-0.7655	0.6444	**0.8421**
	τ	-0.5434	0.5434	0.5385	-0.5448	0.4456	**0.6477**
LOOV-VOC	ρ	-0.7833	0.7833	0.7869	-0.7971	0.6203	**0.8218**
	τ	-0.575	0.575	0.5694	-0.5827	0.4338	**0.6211**
LOOV-COCO	ρ	-0.7941	0.7941	0.7851	-0.785	0.6478	**0.8404**
	τ	-0.5842	0.5842	0.5713	-0.5739	0.4559	**0.6456**
LOOV-LFW	ρ	-0.7551	0.7551	0.7137	-0.7358	0.7032	**0.8462**
	τ	-0.5243	0.5243	0.4683	-0.5001	0.4536	**0.6495**
LOOV-CelebA	ρ	-0.7157	0.7157	0.7082	-0.7542	0.679	**0.8250**
	τ	-0.4875	0.4875	0.4753	-0.5354	0.4569	**0.6270**
Task-Person	ρ	-0.7757	0.7757	0.7833	-0.7997	0.6408	**0.8320**
	τ	-0.5647	0.5647	0.5668	-0.5872	0.4477	**0.6328**
Task-Face	ρ	-0.7435	0.7435	0.6956	-0.7623	0.6756	**0.8590**
	τ	-0.5154	0.5154	0.4537	-0.5387	0.4454	**0.6675**

Table 2. Correlation of various metrics with human anonymity mean scores on $HA2$, following the same evaluation setup as Table 1

ResNet50 backbone architecture for all our setups, trained using Stochastic Gradient Descent (SGD) with a momentum of 0.9 and a learning rate of 0.01. We resize images to 224×224 as expected by these architectures.

5.3. Evaluating Alignment with Human Scores

Results for $HA1$ and $HA2$ correlations across all train and test setups are presented in Tables 1 and 2 respectively, where PerceptAnon, is trained for Regression (mean) – regression on respective HA annotations using mean scores. PerceptAnon consistently has the highest alignment to human scores for anonymity under all training and testing setups and across both $HA1$ and $HA2$ as compared to the image quality and similarity metrics. The 'All' train/test setup gives the highest correlations but in most cases this increase over other setups is not significant. Furthermore, the LOOV experiments reinforce the robustness of PerceptAnon against data distribution shifts. When trained on different datasets, PerceptAnon maintains its high correlation with human perception, highlighting its adaptability.

In the $HA2$ setup, where humans assess anonymity by directly comparing original and anonymized images side by side, perceptual differences play a crucial role. Among conventional metrics, LPIPS and SSIM exhibit the strongest correlation with human scores, likely due to their emphasis on perceptual similarity. However, while these metrics effectively measure perceptual differences, they fall short in identifying specific image regions crucial for maintaining anonymity. In contrast, we believe PerceptAnon, is able to not only understand the degree but also the critical areas of information that remain unobfuscated.

It is essential to note that in $HA1$, PerceptAnon predictions are based solely on the anonymized image, without the original for comparison (which is also the case with $HA1$ annotations themselves). The existing metrics typically rely on direct comparisons with an original image. Since PerceptAnon, when trained on $HA1$, only utilizes the anonymized image, it makes anonymity judgments independently of

the original image, possibly capturing semantic details that humans might use to assess anonymity. This ability to discern privacy-sensitive elements without needing a reference highlights the potential of PerceptAnon in evaluating and understanding image anonymization.

Figure 6 shows that PerceptAnon is able to identify privacy-compromising cues in image backgrounds. We notice that PerceptAnon is not predicting scores based on arbitrary local regions, but considers important regions related to the subject or background as human vision would. This includes notable background regions, auxiliary items/regions related to subjects, and residual details. In Appendix D we include additional heatmaps scenarios to showcase PerceptAnon's focus on individual items or notable background identifiers. We notice that PerceptAnon focuses on relevant background cues of various levels and types across our diverse dataset.

Figure 6. Sample GRAD-CAM visualizations using PerceptAnon (10-class) classification model trained on 'all' train/test setup ($HA1$). More examples are provided in Appendix D.

5.4. Discussions

Regression vs. Classification Strategy. When learning image anonymization through human evaluation, we face a pivotal decision: should this be learned as a regression or a classification problem? While regression may be able to capture the subtleties and variations in human perception with its continuous output, classification simplifies this complexity into broader, more generalizable categories. Our results suggest that the classification strategy tends to be more effective in learning anonymity, particularly evident in Kendall's correlation, as shown in Figure 7. Corresponding Spearman's results can be seen in Figure 10.

Class Granularity in Quantifying Anonymization. In our study, illustrated in Figure 7, we evaluate the effects of learning with different levels of score granularity on anonymity assessments. Our findings reveal that finer granularities, such as 10, 5, or 3 class setups, exhibit significantly higher correlation with human perception compared to the binary approach. We believe human perception of anonymity encompasses a range of nuances, which are more effectively captured by models that operate beyond a simple binary

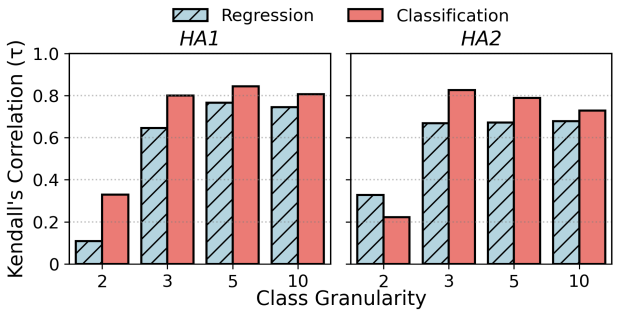


Figure 7. Correlation to human scores across different granularities of anonymization scores (10-class, 5-class, 3-class, binary) and training methods (Regression vs. Classification) for PerceptAnon. Results are based on the 'All' train/test setup. Corresponding Spearman's results are in Appendix C (Figure 10).

classification. This distinctly suggests that from a human perspective, anonymity is not a binary concept but rather a spectrum. In the case of $HA2$, where there is less uncertainty with reference image, we noticed a trend where higher granularity levels tend to yield marginally reduced Kendall's correlation, but Spearman's is constant.

Correlation between $HA1$ and $HA2$. We used common anonymization methods (inpaint and blur/pixelate) on the same image sets to compare alignment of PerceptAnon trained on $HA1$ and $HA2$ (see Table 3). $HA2$-trained PerceptAnon and $HA2$-trained PerceptAnon use the same set of images, with the only distinction being the annotation process. We believe that $HA2$-trained PerceptAnon judgments are influenced by the original image context, whereas $HA1$-trained PerceptAnon focused on detecting any remaining information in the anonymized versions. Since $HA1$ and $HA2$ are not perfectly correlated, their differing perspectives highlight the importance of evaluating under both settings to comprehensively assess anonymization. Therefore, we encourage researchers to adopt both approaches in evaluating anonymized images.

Model	$HA1$	$HA2$	$\rho(HA1, HA2)$
ResNet18	0.8752	0.8345	0.9050
ResNet50	0.8912	0.8346	0.8848

Table 3. Correlation of PerceptAnon's predictions with human scores for $HA1$ and $HA2$, assessed using Spearman's (ρ) on the same image sets treated with inpaint and blur/pixelation. $\rho(HA1, HA2)$ indicates the correlation between PerceptAnon's predictions when trained on $HA1$ and $HA2$.

Global vs Local. Figure 4 indicates that humans typically assign lower scores to global images, suggesting they consider backgrounds in images as containing information that affect perceived anonymity. PerceptAnon was trained on both global and local images—the latter having higher scores due to the absence of backgrounds. We hypothesize

that this likely enables PerceptAnon to mimic human vision more closely, encapsulating the relevance of background elements in anonymity assessments. Hence, PerceptAnon potentially gains a more holistic understanding of what influences anonymity perception, similar to human evaluators.

Expanding the Scope of Anonymization. The focus of current anonymization efforts predominantly lies in obscuring faces and bodies, often extending to PII like license plates or text, typically aligning with pseudonymization objectives. However, the broader context of an image, including background elements and personal items, can also inadvertently reveal identities, such as the interior of a home or specific personal belongings. Recognizing a familiar setting or item can lead to de-anonymization, despite the anonymization of direct identifiers. Future research should broaden its scope to include these wider aspects, investigating anonymization techniques effective across various settings, both public and private. PerceptAnon represents an initial step in this direction by introducing a metric designed to align with a global human anonymity viewpoint.

Computer Vision vs. Human Vision. In the contemporary digital landscape, images are interpreted by both computer vision (CV) and humans, each with distinct capabilities. Our metric is designed to be used alongside traditional CV-based detectability metrics, expanding the assessment to encompass privacy across the entire image content. While detecting individuals is a critical aspect, understanding the broader privacy implications within the entire image is equally important. PerceptAnon addresses this by evaluating privacy-compromising cues remaining within anonymized images. This comprehensive assessment approach aligns more closely with how humans perceive and interact with anonymized images, thereby providing a more thorough understanding of the effectiveness of anonymization techniques.

Sensitivity to Privacy-Compromising Cues. We showcase a broad selection of anonymized images and PerceptAnon's activation maps in Figure 6 and Appendix D. These examples visually show the cues PerceptAnon detects and overlooks. Our dataset was curated to include a diverse array of images featuring various privacy cues, with train/test splits designed to study identification of cues at different scales. For instance, LOOV scenarios with full-body images focus on large-scale cues such as bicycles, whereas face-only tests concentrate on finer cues like clothing accessories or distinct backgrounds. Nonetheless, thoroughly analyzing subtle privacy-compromising cues remains a complex challenge due to the nuanced aspects of human vision and privacy evaluation. PerceptAnon lays a foundational groundwork for future in-depth exploration and refinement in the field.

New/unseen Anonymization Methods. GDPR defines pseudonimization as processing data such that the subject is unidentifiable, with masking most feasible for images (Weitzenboeck et al., 2022; Barta, 2018). Our study focuses on prevalent anonymization techniques in current research and practical use (Hukkelås & Lindseth, 2023b; Pachni A., 2022). Existing methods typically focus on local regions – faces and bodies. In contrast, PerceptAnon evaluates privacy leakages throughout the entire image, including residual objects and background cues. We hypothesize that PerceptAnon effectively assesses privacy leakages beyond these local methods, hence adapt to unseen anonymization methods that continue to act on local regions. Further experimental results supporting this are detailed in Appendix C.

Limitations. PerceptAnon, while a step forward in human-centric anonymization assessment, can encounter challenges in accommodating diverse cultural perceptions of privacy and managing images of varied complexity. Further exploration into contextual understanding of images and refining the model's ability to identify subtle cues and background elements is important for anonymity assessment. Furthermore, one improvement can be by enhancing the dataset with more diverse cultural contexts and encapsulate more range in scores. Figure 4 reveals a clustering of human anonymity scores towards the higher end, suggesting perceptions of partial anonymity. While we experimented using weighted CE loss to accodomate for this (see Figure 11), there were no substantial performance increases.

6. Conclusion

Through this research, we shed light on the pivotal role of human perception in image anonymization–a facet absent in prior studies. We began by anonymizing images and subsequently collecting human scores to assess the perceived anonymity of the entire image, diverging from the pseudonymization-focused studies that mainly concentrate on re-ID rates in regions with removed PII. To assess (global) anonymity, we utilize traditional image assessment metrics such as SSIM and LPIPS. Our findings reveal that these metrics, not originally designed to capture subjective differences, do not robustly align with the human perspective on image anonymity. Hence, we introduced PerceptAnon, a learning-based metric which better aligns with human perspective, thus addressing concerns raised by regulations like the General Data Protection Regulation (GDPR). PerceptAnon distinguishes itself by not only comparing original-anonymized pairs, but also its ability to evaluate the anonymity of images without original counterpart, mirroring real-world scenarios where individuals encounter anonymized content devoid of reference images. In future work, our focus will be on expanding PerceptAnon's applicability and robustness, expanding our datatset to improve its generalizability and effectiveness in anonymity assessment.

Impact Statement

This study advances image anonymization by introducing a metric attuned to human perceptions of anonymity. We hope this research encourages future efforts to enhance privacy protection in digital media, considering entire images and guides the ethical use and development of anonymization technologies.

References

Barattin, S., Tzelepis, C., Patras, I., and Sebe, N. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8001–8010, 2023.

Barta, G. Challenges in the compliance with the general data protection regulation: anonymization of personally identifiable information and related information security concerns. *Knowledge–economy–society: business, finance and technology as protection and support for society*, pp. 115–121, 2018.

Cao, F., Sun, J., Luo, X., Qin, C., and Chang, C.-C. Privacy-preserving inpainting for outsourced image. *International Journal of Distributed Sensor Networks*, 17(11): 15501477211059092, 2021.

Dietlmeier, J., Antony, J., McGuinness, K., and O'Connor, N. E. How important are faces for person re-identification? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6912–6919. IEEE, 2021.

Du, L., Zhang, W., Fu, H., Ren, W., and Zhang, X. An efficient privacy protection scheme for data security in video surveillance. *Journal of visual communication and image representation*, 59:347–362, 2019.

EDPS. Guidelines on anonymisation: Minsunderstandings related to anonymisation. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf, 2021.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Fredrik Lundh, J. A. C. Pillow ImageFilter Module. https://pillow.readthedocs.io/en/stable/reference/ImageFilter.html, 2024. Accessed: Jan 2024.

Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., and Isola, P. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pp. 505–520. Springer, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

He, X., Zhu, M., Chen, D., Wang, N., and Gao, X. Diff-privacy: Diffusion-based face privacy protection. *arXiv preprint arXiv:2309.05330*, 2023.

Hellmann, F., Mertes, S., Benouis, M., Hustinx, A., Hsieh, T.-C., Conati, C., Krawitz, P., and André, E. Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *arXiv preprint arXiv:2305.02143*, 2023.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Hukkelås, H. and Lindseth, F. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1329–1338, 2023a.

Hukkelås, H. and Lindseth, F. Does image anonymization impact computer vision training? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 140–150, 2023b.

Hukkelås, H., Mester, R., and Lindseth, F. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pp. 565–578. Springer, 2019.

Hukkelås, H., Smebye, M., Mester, R., and Lindseth, F. Realistic full-body anonymization with surface-guided gans.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1430–1440, 2023.

Jaichuen, T., Ren, N., Wongapinya, P., and Fugkeaw, S. Blur & track: Real-time face detection with immediate blurring and efficient tracking. In *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 167–172. IEEE, 2023.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., and Huang, F. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5060–5069, 2019.

Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., and Cao, X. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3891–3899, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Maximov, M., Elezi, I., and Leal-Taixé, L. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5447–5456, 2020.

McPherson, R., Shokri, R., and Shmatikov, V. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

Nagrani, A., Albanie, S., and Zisserman, A. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8427–8436, 2018.

Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. Faceless person recognition: Privacy implications in social media. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pp. 19–35. Springer, 2016.

Pachni A. Blur faces in videos automatically with amazon rekognition video. https://aws.amazon.com/blogs/machine-learning/blur-faces-in-videos-automatically-with-amazon-rekognition-video/, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Peng, C., Wan, S., Miao, Z., Liu, D., Zheng, Y., and Wang, N. Anonym-recognizer: Relationship-preserving face anonymization and recognition. In *Proceedings of the 3rd International Workshop on Human-Centric Multimedia Analysis*, pp. 1–6, 2022.

Qian, J., Li, X.-Y., Zhang, C., Chen, L., Jung, T., and Han, J. Social network de-anonymization and privacy inference with knowledge graph model. *IEEE Transactions on Dependable and Secure Computing*, 16(4):679–692, 2017.

Rosenberg, H., Ahmed, S., Ramesh, G. V., Vinayak, R. K., and Fawaz, K. Unbiased face synthesis with diffusion models: Are we there yet? *arXiv preprint arXiv:2309.07277*, 2023.

Saunders, B., Kitzinger, J., and Kitzinger, C. Anonymising interview data: Challenges and compromise in practice. *Qualitative research*, 15(5):616–632, 2015.

Shao, Y., Liu, J., Shi, S., Zhang, Y., and Cui, B. Fast de-anonymization of social networks with structural information. *Data Science and Engineering*, 4:76–92, 2019.

Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5179–5187, 2015.

Sun, X., Gazagnadou, N., Sharma, V., Lyu, L., Li, H., and Zheng, L. Privacy assessment on reconstructed images: Are existing evaluation metrics faithful to human perception? *arXiv preprint arXiv:2309.13038*, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Upenik, E., Akyazi, P., Tuzmen, M., and Ebrahimi, T. Inpainting in omnidirectional images for privacy protection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2487–2491. IEEE, 2019.

Vishwamitra, N., Knijnenburg, B., Hu, H., Kelly Caine, Y. P., et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 39–47, 2017.

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

Wiles, R., Coffey, A., Robinson, J., and Heath, S. Anonymisation and visual images: issues of respect,'voice'and protection. *International Journal of Social Research Methodology*, 15(1):41–53, 2012.

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

Yang, S., Luo, P., Loy, C.-C., and Tang, X. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019.

Yuan, L., Liu, L., Pu, X., Li, Z., Li, H., and Gao, X. Pro-face: A generic framework for privacy-preserving recognizable obfuscation of face images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1661–1669, 2022.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., and Li, H. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2737–2746, 2020.

Zoom Support. Zoom: Using blurred background. https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0061066, 2023.

A. Dataset Examples

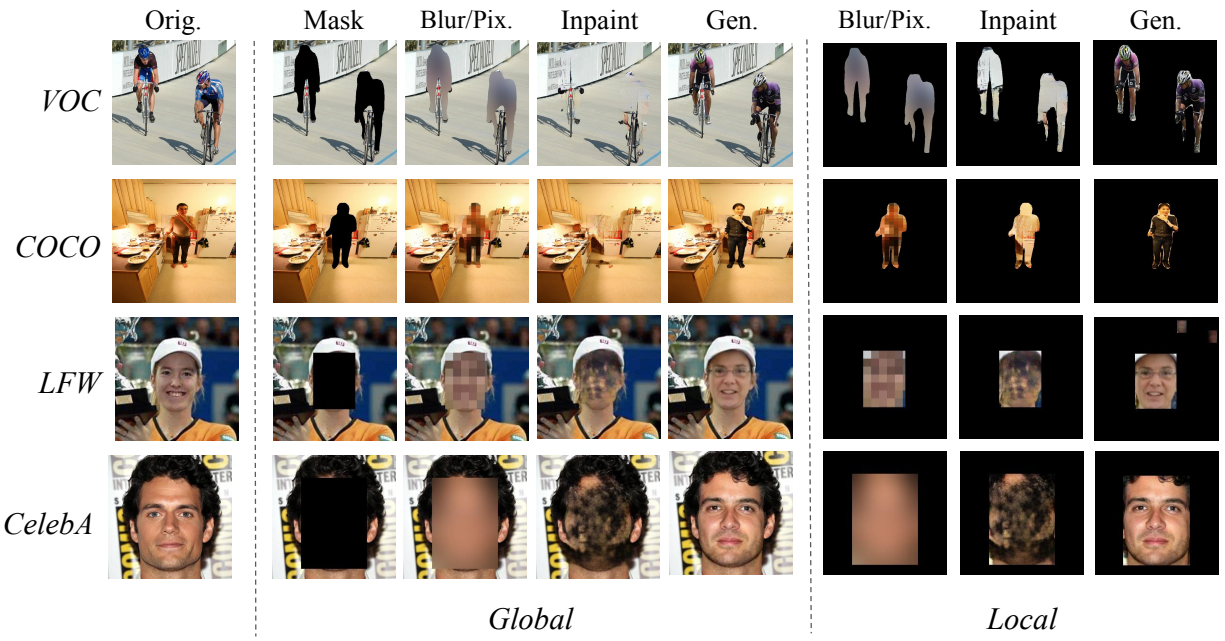


Figure 8. Example images from our curated dataset where we specifically anonymize PII; anonymizing faces in the LFW and CelebA datasets and full bodies in the COCO and VOC datasets. In each image, the PII is anonymized using traditional methods including masking, blur/pixelation and deep learning-based methods including inpainting (removal) and generative method. *Global* is akin to considering true anonymization and *Local* to pseudonymization.

B. Training and Model Details

For all model training, we utilized minor augmentations, including RandomHorizontalFlip and RandomRotation. We deliberately avoided stronger augmentations, such as cutout, to prevent potential confusion in the models or loss of crucial background details necessary for understanding. For instance, employing cutout might lead the model to mistakenly interpret the cutout region as being anonymized. The models were trained over 200 epochs, and we selected the model with the lowest validation loss for subsequent use.

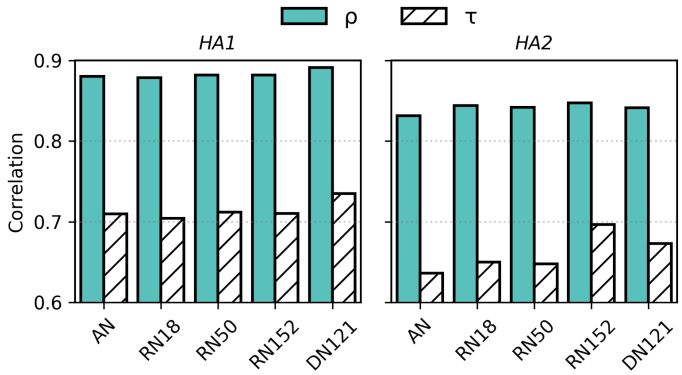


Figure 9. Evaluating the impact of different backbone architectures on PerceptAnon. AN, RN, and DN denote AlexNet, ResNet, and DenseNet, respectively. Both Spearman's (ρ) and Kendall's (τ) correlation results against human scores are shown. Results are on the 'All' train/test split using mean human scores (regression).

Backbone Architectures. We trained PerceptAnon with different backbones, Figure 9 shows the results on the 'All' Train/Test setup. We notices there is no clear trend or significant difference. Notably, using deeper architectures like ResNet152 and DenseNet121 gave higher Kendall's correlation results on $HA2$.

Choice of CNNs. We initially explored metric learning.We employed backbone architectures and metric learning strategies similar to those used in SemSim (Sun et al., 2023), specifically utilizing triplet loss and batch hard negative mining to compute the L2 distance between embeddings of original and anonymized images. However, strong CNN performance in early tests led us to adopt these networks as our primary methodology. The robustness of CNNs facilitated a straightforward and interpretable approach, crucial for our objective of evaluating models. Our study does not primarily focus on architectural choices but rather on an exploration of diverse annotation setups and the comprehensive analysis of anonymization methods applied to faces and personal identifiers.

C. Additional Results

Unseen Anonymization Methods. Established techniques such as masking, blurring, and generating are the predominant in current research and practical applications (Hukkelås & Lindseth, 2023b; Zoom Support, 2023; Yang et al., 2022). While our work aims to establish a foundational understanding of image anonymity using these well-recognized methods, we additionally evaluate how PerceptAnon would fare when faced with new or unseen anonymization methods. We employ another leave-one-out-validation (LOOV) strategy, similar to Section 5.1 and Table 1, where one anonymization method is excluded from the training set and introduced only in the test set. For instance, in the LOOV-Mask scenario, we train on inpainted and blur/pixelated images, and test on masked images. Results shown in Table 4 indicate that PerceptAnon maintains robust performance but struggles with masked images, as masking removes all underlying information, unlike blurring or pixelation which preserves more context. This highlights PerceptAnon's potential utility with unseen anonymization techniques, although its performance can vary based on the nature of the anonymization method used. Future work can build on PerceptAnon's adaptability to new methods by expanding our dataset to include them.

Train/Test Setup	Metrics	PSNR	MSE	LPIPS	SSIM	FID	PerceptAnon (Ours)
LOOV-BlurPix	ρ	-0.8191	0.8191	0.8106	-0.8071	0.6577	**0.8862**
	τ	-0.6167	0.6167	0.6053	-0.6020	0.4638	**0.7192**
LOOV-Inpaint	ρ	-0.7947	0.7947	0.8004	-0.7791	0.7088	**0.8646**
	τ	-0.5850	0.5850	0.5927	-0.5615	0.5181	**0.6907**
LOOV-Mask	ρ	-0.1780	0.1780	0.2703	-0.2957	0.2949	**0.5816**
	τ	-0.1240	0.1240	0.1926	-0.2125	0.2055	**0.4275**

Table 4. Correlation of various metrics with human anonymity mean score – Regression (mean) on $HA1$, assessed using Spearman's (ρ) and Kendall's (τ) correlations. Results are for Leave-One-Out Validation (LOOV) on anonymization methods where LOOV-Mask indicates using Masking as the test dataset and remaining as train.

Regression vs. Classification, Class Granularity Figure 10 shows Spearman's correlation results for varying class granularities and regression vs. classification setups (correspondent results are in Figure 7. Similar to Kendall's, Spearman correlation results also favor classification over regression, but less significantly than Kendall's. Choosing a binary setup is again noticably least performant, but using Spearman's correlation, the changes in granularities of 3, 5, and 10 are less significant compared to Kendall's correlation. Particularly in the case of $HA1$ in figure 10, where there is more uncertaintly without reference image, we noticed a trend where higher granularity levels tend to yield marginally better performance.

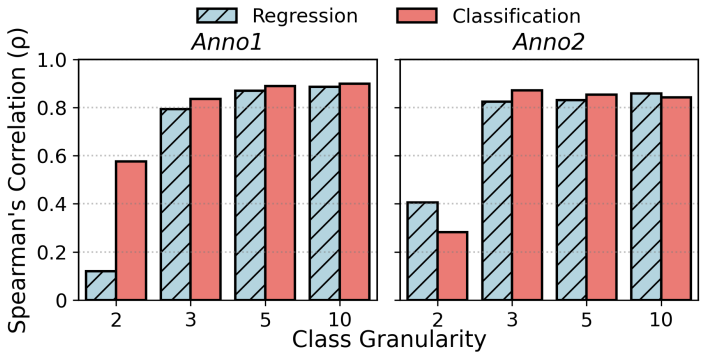
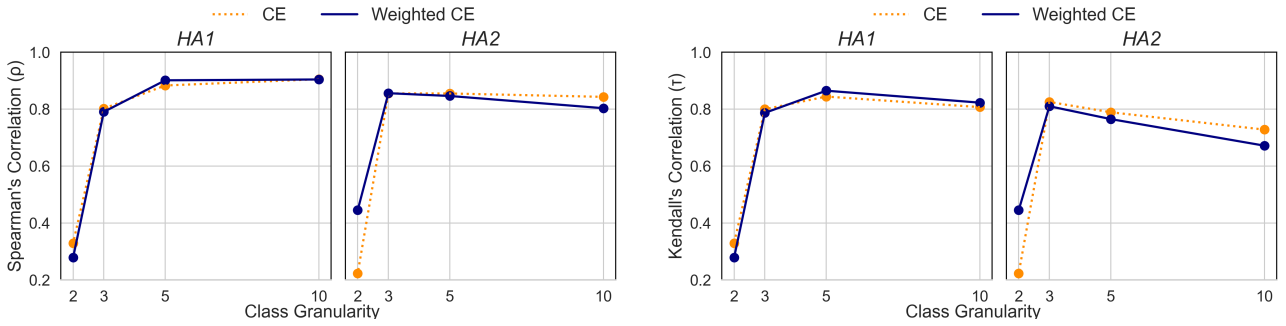


Figure 10. Corresponding Spearman's correlation results to Figure 7. Correlation to human scores across different granularities of anonymization scores (10-class, 5-class, 3-class, binary) and training methods (Regression vs. Classification) for PerceptAnon. Results are based on the 'All' train/test setup.

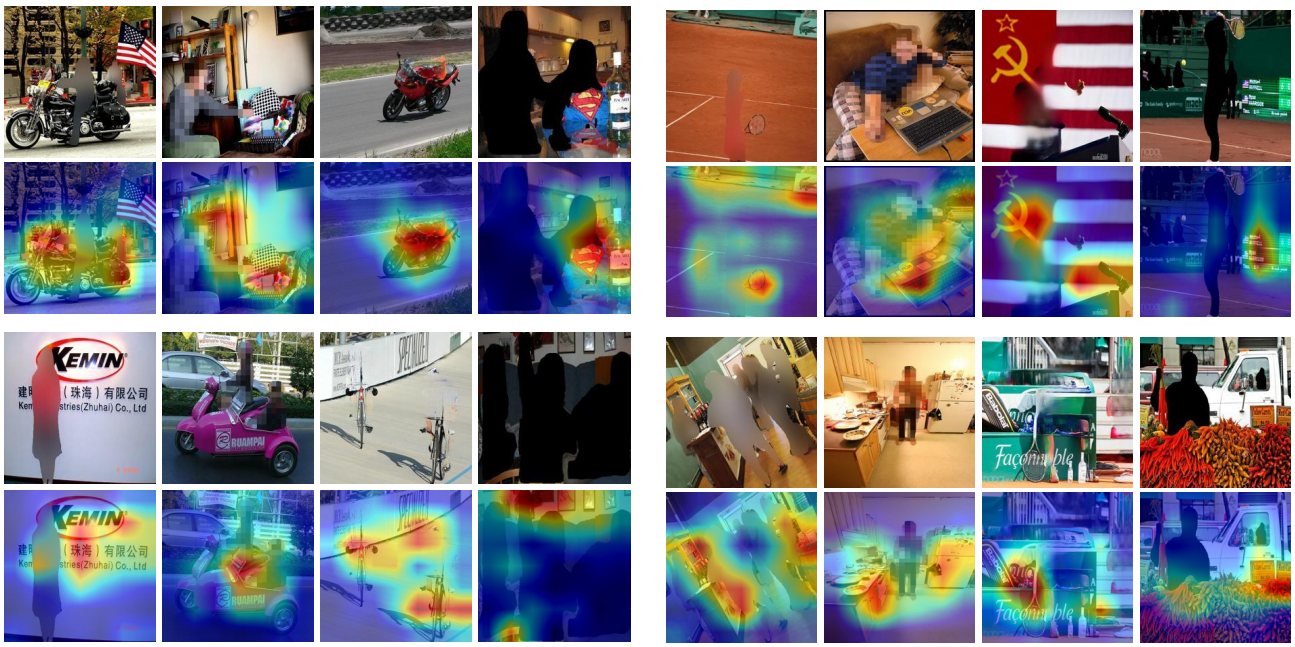
Using Weighted Cross-Entropy Due to the score imbalance observed in Figure 4, where the majority of images are scored above 5 (on a scale from 0 to 10, with 0 being the lowest and 10 the highest in terms of anonymity), we experimented with using weighted Cross-Entropy (CE) in our classification strategy for the 'All' test/train setup. The comparative results between standard CE and Weighted CE are presented in Figure 11. On $HA1$, the trends are consistent for both Spearman's and Kendall's correlations, showing a marginal increase in correlation. However, this trend does not extend to $HA2$, where we observe a reduced correlation when using weighted CE.



(a)) Spearman's Correlation (b)) Kendall's Correlation

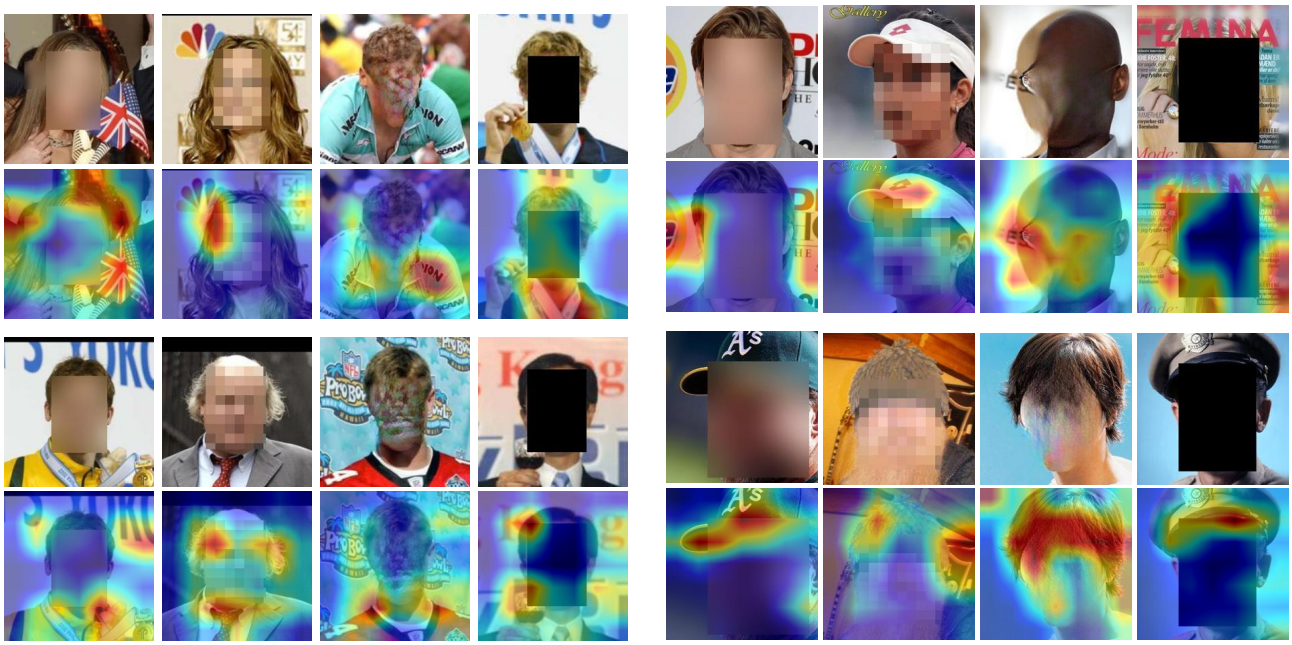
Figure 11. Performance of Cross-Entropy (CE) vs. Weighted Cross-Entropy loss (Weighted CE) for classification on different class granularities of anonymization scores. Results are on the 'All' train/test setup.

D. PerceptAnon Activation Maps



(a)) VOC (b)) COCO

Figure 12. Sample GRAD-CAM visualizations from our dataset using PerceptAnon (10-class) classification model trained on 'all' train/test setup ($HA1$). We showcase each source dataset (VOC, COCO, LFW, CelebA) images seperately.



(a)) LFW (b)) CelebA

Figure 13. Sample GRAD-CAM visualizations using PerceptAnon from our dataset for source datasets LFW, CelebA.

E. Anonymization Methods

Before anonymization, masks are generated with Dual Shot Face Detector (DSFD) (Li et al., 2019) for faces and Mask R-CNN (He et al., 2017) for full body segmentation, chosen for its minimal impact on the surrounding areas. These masks guide targeted anonymization.

E.1. Masking

Masking or Mask-out involves overlaying a specific region in an image (such as a face or full body) with a mask, which could be a solid color or a pattern.

Mathematical Formulation: Given an RGB image I and a mask M (where $M(x, y) = 1$ for pixels to be masked and 0 otherwise), the anonymized (masked) image I' is given by:

$$I' = I \odot (1 - M) + M \odot C \tag{1}$$

where $\odot$ represents elementwise multiplication, and $\mathbf{C}$ is a vector representing the RGB color used for masking. In this work, black color is used, so $C = (0, 0, 0)$.

E.2. Blur & Pixelation

Both blurring and pixelation are applied to specific regions of an image as defined by masks. This process involves extracting the regions indicated by the masks, applying the anonymization technique (blurring or pixelation), and then integrating these transformed regions back into the original image. The general process can be defined below:

1. For each mask M_i defining the region to anonymize, extract the corresponding region from the original image I.
2. Apply the desired transformation (blurring or pixelation) to the extracted region.
3. Replace the original region in image I with the transformed region to obtain the anonymized image.

Mathematical Formulation: Let I be the input image, M_i be a mask, and T be the transformation function (either blurring or pixelation). The anonymized image I' is obtained by:

$$I' = (1 - M_i) \odot I + M_i \odot T(I_{M_i}) \tag{2}$$

where I_{M_i} is the region of I defined by mask M_i, and $T(I_{M_i})$ is the transformed region.

E.2.1. Gaussian Blur Transformation

The Gaussian blur transformation T_{blur} on a region I_{M_i} is defined as:

$$T_{\text{blur}}(I_{M_i}) = \text{GaussianBlur}(I_{M_i}, \sigma) \tag{3}$$

where $\text{GaussianBlur}(\cdot, \sigma)$ applies Gaussian blur with standard deviation σ to the region. We utilize PIL's ImageFilter.GaussianBlur() function (Fredrik Lundh, 2024) and use filter radius equal to 1/8 of the width of the bounding box, as (Dietlmeier et al., 2021) showed it effectively removes all the identifying features.

E.2.2. Pixelation Transformation

The pixelation transformation T_{pixelate} on a region I_{M_i} is defined as:

$$T_{\text{pixelate}}(I_{M_i}) = \text{Pixelate}(I_{M_i}, \text{block\_size}) \tag{4}$$

where $\text{Pixelate}(\cdot, \text{block\_size})$ applies pixelation with a block size of 16×16, in line with (Hukkelås et al., 2019).

E.3. Inpainting & Generative

For both inpaint removal and generative methods, we use the same face detectors and segmentation model as traditional method.

Inpaint (removal) We utilize DeepFillV2 implementation based off (Yu et al., 2019), which uses gated convolutions in a coarse-to-fine manner.

Generative We use DeepPrivacy2 (Hukkelås & Lindseth, 2023a) which offers both face and fulll body anonymization. It is the first work to target full body generative anonymization. This work utilizes dense pose estimation and a style-based GAN.

F. Image Assessment Metrics

Each metric has unique characteristics in terms of how it interprets image similarity or dissimilarity, which is crucial for understanding its implications for image anonymity.

Mean Squared Error (MSE): MSE measures the average squared difference between pixels of two images. A lower MSE value indicates greater similarity between the images. In the context of anonymity, a higher MSE is desirable as it suggests that the anonymized image significantly differs from the original.

Peak Signal-to-Noise Ratio (PSNR): PSNR is a pixel-level measure of the peak error between two images. Similar to MSE, a lower PSNR indicates more similarity, and thus, for anonymity, a higher PSNR value is preferable.

Structural Similarity Index (SSIM): SSIM evaluates the visual impact of changes in luminance, contrast, and structure between two images. SSIM values range from -1 to 1, with higher values indicating more similarity. Therefore, in terms of anonymity, lower SSIM values are better as they imply greater dissimilarity between the original and anonymized images.

Learned Perceptual Image Patch Similarity (LPIPS): LPIPS uses deep learning, specifically AlexNet in our evaluation, to estimate perceptual similarity. A lower LPIPS value indicates higher perceptual similarity. For anonymity, a higher LPIPS score is desired, suggesting that the anonymized image is perceptually distinct from the original.

Fréchet Inception Distance (FID): FID assesses the similarity in the distribution of features extracted by a model from two sets of images. IT is typicaly used to assess the quality of images generated by generative adversarial networks (GANs) (Goodfellow et al., 2014). A lower FID indicates closer feature distributions, implying similarity. For anonymity purposes, a higher FID score is preferred, indicating a greater dissimilarity in feature distribution between the anonymized and original images.