Date of publication: 26-03-2025

Computer Vision with Anonymized Data: A Systematic Approach for Evaluation using Realistic Anonymization

SARAH WEISS[1], CHRISTOPHER BONENBERGER[2], MARKUS SCHNEIDER[2]
[1]UNIVERSITY OF APPLIED SCIENCES at Ravensburg-Weingarten, (Web: www.rwu.de, Mail: weisssa@rwu.de, Matr.: 35655)
[2]UNIVERSITY OF APPLIED SCIENCES at Ravensburg-Weingarten

ABSTRACT

Camera-based systems in Ambient Assisted Living (AAL) and Autonomous Driving (AD) require careful handling of privacy-sensitive image data. The ideal way to prevent data misuse is to anonymize data right after perception and before processing. Non-realistic anonymization methods (blur, pixelation) suffice, but remove essential information needed by subsequent algorithms. Realistic anonymization, on the other hand, promises to preserve vital information, by generation of natural-like replacements. Recent studies investigate the performance on such data but do not examine the underlying causes of the observed impacts. For that reason, this study aims to establish a systematic approach to analyze anonymization methods and their effects on model training and performance, through a quantitative review of the challenges and changes introduced by anonymization.

By using the state-of-the-art toolbox DeepPrivacy2, we generate a realistic full-body anonymized COCO dataset and use it to train and evaluate YOLOv10 on object detection. In addition to classic metrics (mAP, AP), the Structural Similarity Index Measure (SSIM) is utilized to assess the impact of anonymization on images or classes. To gain insights on the influences of anonymization on computer vision, we conduct experiments focusing on factors like object size, as well as co-occurrence frequency with the anonymized class 'person'. Furthermore, novel findings on the robustness of model sizes and the processing of anonymized images within the model are presented.

Training and evaluation with anonymized data pose challenges like object obfuscation and re-labeling. Results indicate that future research must adapt models to anonymized data, improve realistic anonymization generation, and provide datasets suited for research in anonymization. This will help establish life-changing technologies like AAL and AD and narrow the gap between privacy and the information demands of computer vision.

INDEX TERMS

Anonymization, Computer Vision, DeepPrivacy2, Object Detection, Training Data, YOLO

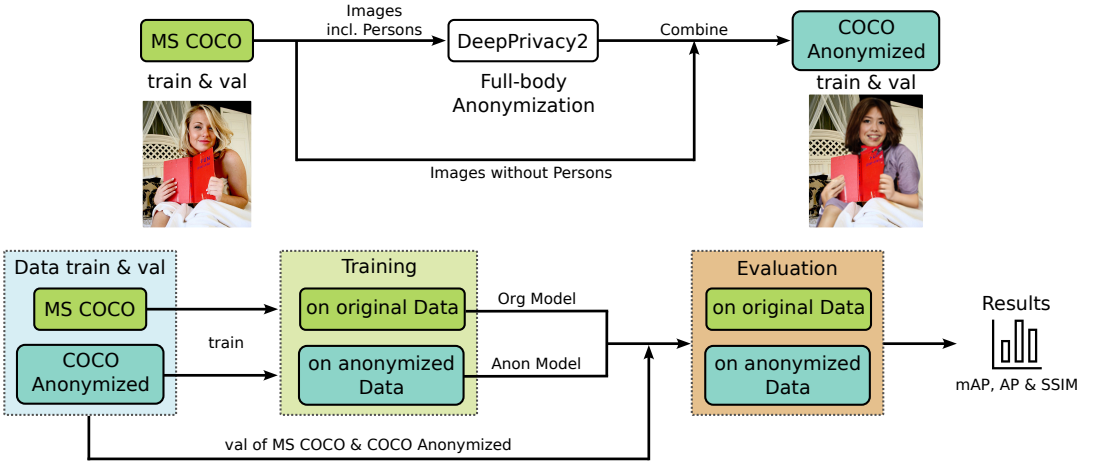

FIGURE 1. Top: Generation of an anonymized COCO dataset using DeepPrivacy2 on the full training and validation set of COCO. Bottom: Visualizing the data flow for training and evaluation. Usage of either the original COCO data or anonymized COCO for training results in models specialized for the respective data type (Org Model, Anon Model). Each model type is evaluated both on original and on anonymized data. The full training and validation set is used, without filtering for hard or complex scenes. Evaluation parameter are mAP, AP and SSIM.

1 INTRODUCTION

Ambient Assisted Living (AAL) and Autonomous Driving (AD) are emerging fields with significant potential to enhance the quality of life. AAL empowers elderly adults and individuals with physical or mental impairments to manage daily tasks more independently — a key advantage as care facilities face surging demand from demographic shifts and a simultaneous caregiver shortage [1]. AD also has the potential to generate societal benefits by improving mobility access for individuals who are unable to drive. As well as environmental benefits through more efficient driving [2].

Both in AAL and AD, key technologies like the Internet of Things (IoT), data fusion, Artificial Intelligence (AI), and cloud computing offer new possibilities. Nevertheless, they also raise security and privacy concerns due to an increasing number of sensors monitoring everything in their field of view. From a technical perspective, this level of monitoring is intentional, but introduces risks of data abuse. Incidents of AD in recent years highlight these risks, such as Tesla employees sharing sensitive footage from customer cars [3] and misuse of GPS trackers in cars [4].

As applications for AAL become more widespread, the potential for data misuse will likely increase. This emphasizes the growing importance of privacy-preserving technologies and anonymization to mitigate these risks or mitigate the consequences. In AAL, high levels of trust are essential as the technology requires continuous monitoring in personal spaces, revealing private details such as habits, health data, personal hygiene, and sexual preferences. These privacy risks extend beyond the intended users, like patients or medical staff, to include bystanders, such as visitors or service personnel [5]. Even with declared data confidentiality, there are no guarantees for privacy — especially with cloud-based processing.

The issue of bystanders is equally relevant for AD, as future vehicles equipped with cameras, LiDAR, and other sensors will continuously survey both the public environment and the driver. While drivers — or in AAL contexts, patients and caregivers — may consent to monitoring, the consent of other individuals in public spaces remains questionable. For pedestrians, potential data leaks heighten the risk of being constantly trackable, especially when data is aggregated and analyzed on a large scale.

This is not only an ethical issue, but also a matter of legal compliance. Since 2018, the European General Data Protection Regulation (GDPR) [6], one of the strictest data protection laws, mandates explicit individual consent for processing personal data. As a result, anonymization is critical not only at the application layer, but also requires advancement in scientific research.

Training AI requires vast amounts of data, and creating new datasets is costly and time-consuming. A critical challenge arises when individuals withdraw consent for data usage. Such actions can potentially render entire datasets worthless, posing a significant challenge to AI development.

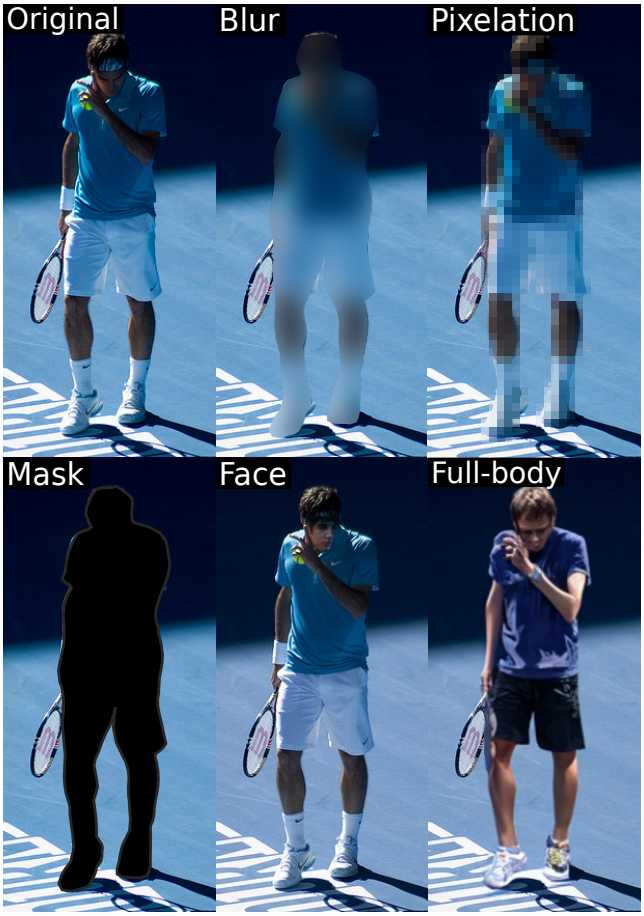

FIGURE 2. Examples for different methods of anonymization using the same base image. Showing non-realistic methods like blur, pixelation and masking as well as realistic methods through DeepPrivacy2 with face and full-body anonymization. It can be seen how different methods can cause loss of objects within the anonymized region, e.g. the tennis ball and the wristband.

Anonymization offers a possible solution by allowing continued use of existing data, which raises questions about whether models trained on anonymized data can still perform effectively on original inputs.

If existing models remain suitable, no additional fine-tuning on anonymized data is required, and resources are saved; otherwise, new models specialized for anonymized data may be necessary. This has practical implications because not every application has the resources to train or maintain multiple models, so a single model that performs well on both original and anonymized data is ideal.

The challenges outlined underscore the need for anonymization methods with robust privacy guarantees. These methods must ensure no decline in performance on the intended application and, therefore, must still preserve required information.

Classic non-realistic anonymization are methods such as pixelation, blurring, or masking as shown in Figure 2. These often rely on transforming the source data or on inpainting. Transformation-based methods face reversibility issues (e.g., shown in [7]), therefore failing to provide strong privacy

guarantees. Both approaches — transformation based and inpainting — also suffer from information loss, potentially rendering further processing harder or impossible, e.g., pose estimation for action recognition on blurred or masked data (compare in Figure 2).

Realistic anonymization offers a promising solution to these issues by replacing information with synthesized content generated by AI. These aim for natural-like replacements. So called full-body anonymization, generates replacements based on poses and provides stronger privacy guarantees than classic or purely transformative methods. A notable recent example in this field is DeepPrivacy2 [8], which is utilized in this work to achieve realistic anonymization, as seen in Figure 2.

While realistic anonymization addresses some issues, it keeps known challenges and introduces new ones. Objects carried by anonymized individuals may be lost during replacement, leading to information loss (compare Figure 2, tennis ball and wristband are lost by anonymization with blur, pixelation, mask and full-body). Additionally, replacements may introduce imperfections, such as changes in the texture of surrounding objects or changes in illumination, potentially degrading the performance of subsequent processing or AI training. These factors highlight the complexity of achieving effective anonymization for use in real-world applications.

In addition, there is a lack of datasets specifically designed to investigate computer vision tasks under the influence of anonymization. This gap shows the lack of research on the effects of anonymization, especially as the volume of data featuring anonymized individuals and objects continues to grow in the future.

Given these challenges, this work examines the influence and consequences of realistic anonymization on AI training. We focus on YOLOv10 [9], the latest version of the most widely used object detection models, at the start of this study. The aim is to obtain a broad overview of common issues related to anonymized data. The key questions are:

- Does training on realistic anonymization improve detection performance compared to original data, and how does it compare to non-realistic methods?
- What influence does model size (parameter count) have on the object detection performance of YOLO in the context of anonymization?
- Do models maintain their detection performance if processed data type (anonymized or original) is exchanged for a different one than the model is trained on?
- How much does realistic anonymization alter the image or specific objects?
- How much does object size and frequency with the anonymized class affect detection performance?
- Is fine-tuning pre-trained models with anonymized data sufficient to improve detections on anonymized inputs?
- Which parts of the detection process are most affected by anonymization changes?
- Is the investment in a specialized dataset for anonymization research worthwhile?

2 RELATED WORK

Research on anonymization in AI training primarily focuses on evaluating anonymization methods or their impact on person detection, with fewer studies examining the effects of anonymized data on model training. Traditional methods like blurring and pixelation often degrade performance, while realistic approaches, such as DeepPrivacy2, offer improved results. However, even realistic anonymization affects computer vision tasks, with the impact varying by model architecture and used anonymization method [10–12]. Broader discussions on the general topic of visual privacy highlight the limitations of traditional anonymization and emphasize the need for more advanced full-body anonymization methods [13]. This section reviews key works addressing these challenges and the role of anonymization in model training.

2.1 ANONYMIZATION METHODS

To better understand the impact of anonymization on model training, it is essential to examine the different anonymization methods available, ranging from traditional techniques to more advanced realistic approaches. The work of [13] provides a broad overview of privacy-preserving techniques, extending beyond anonymization, to explore various methods for visual privacy.

It discusses intervention approaches, data hiding, visual obfuscation techniques such as image filtering, gait anonymization, traditional and realistic anonymization techniques, and poisoning attacks, among others. Additionally, the review dives into privacy-by-design systems with different privacy levels.

The review emphasizes that simply anonymizing facial features is insufficient, as other identifiable attributes, such as gait, gender, and height, can still be extracted. This underlines the importance of realistic full-body approaches for anonymization to address all potential identifiers effectively.

Traditional, non-realistic anonymization methods, such as pixelation and blurring, are ineffective for computer vision tasks due to significant information loss and being reversible [7, 14, 15]. Realistic anonymization methods are more sophisticated due to the generation of new content. But these remain limited, with only a few frameworks focusing on full-body anonymization, including [8, 16–18]. For instance, DeepPrivacy2 exemplifies the latest advancements in realistic anonymization.

DeepPrivacy2

DeepPrivacy2 is a GAN-based toolbox designed for face and full-body anonymization. Its detection step integrates three frameworks, consisting of face detection using DSFD [19], pose estimation via CSE [20], and instance segmentation through Mask R-CNN [21]. Using these detection steps, individuals are grouped into those with poses detected by CSE, persons not identified through CSE, and faces missed by the other two methods. Each group undergoes an own specialized anonymization process.

The anonymized individuals are re-integrated into the original image using a recursive ordering method, designed to minimize stitching artifacts. Larger detections are presumed to be in the foreground. These are stitched last to maintain scene integrity, while smaller ones are placed first.

DeepPrivacy2 includes notable features, such as the ability to track individuals across image sequences and preserve generated identities, ensuring consistency. It can reproduce identical anonymization results when processing the same source image repeatedly, further enhancing consistency. Additionally, generation can be guided by text prompts, enabling control over anonymized attributes — for example, specifying that all generated individuals should have specific expressions like smiling or other facial attributes like having mustaches. Therefore, it offers a wide range of possible applications.

While the authors identify a limitation in how the generator's output is influenced by image context (e.g., generating a baseball player when a baseball field is detected in the surrounding environment), this work highlights it as a potential advantage. Refining this characteristic in future iterations could enable context-aware identity generation, preserving critical information and enhancing realism.

2.2 IMPACT OF ANONYMIZED TRAINING DATA ON COMPUTER VISION

Research on the impact of anonymized data on AI training and its effects on computer vision tasks remains scarce. Among recent studies, three stand out as they move beyond merely assessing whether new anonymization methods still enable person detection or evaluate the level of privacy they provide. Instead, these works examine how training on anonymized data affects object detection, segmentation, and keypoint detection, offering valuable insights into the implications of anonymization on computer vision. As these studies serve as the main inspiration for our work, their findings and methods are summarized in this section, and key differences and contributions to this research are highlighted in Section 2.5.

The study of LEE [10] examines non-realistic anonymization techniques, and their effects on different computer vision tasks when they are used as anonymized training data. Additionally, they compare different architectures, such as ViT-based models and CNN-based models (YOLOv8 [22]). The results indicate that CNN architectures are more robust in regard to performance degradation caused by anonymized data. Further, an interesting finding is that object classes frequently appearing alongside anonymized persons experience a decrease in accuracy, while the impact on other classes is considered negligible.

This suggests that anonymized classes can negatively influence the performance of non-anonymized ones. The extent of this impact depends on the anonymization technique, model architecture, and class type. The study also found that larger models with more parameters are less affected by anonymization-related degradation. However, no information is provided on how objects or labels within anonymized regions are handled. These can pose a problem for evaluation if objects are removed during anonymization.

The impact of face-anonymized training data on segmentation tasks within AD is investigated by ZHOU [11]. They apply different anonymization techniques, including blur, crop-out, random crop, and realistic anonymization. Their study uses the Cityscapes dataset [23], which is specialized for autonomous driving. By focusing on use case relevant classes (persons and three vehicle types) the evaluation is highly use case specific. Their results indicate that, among the applied anonymization methods, realistic face anonymization performs best, but still leads to a decrease in performance. Additionally, like LEE [10], they find that larger models are less affected by anonymization-induced degradation compared to smaller ones.

Another study focusing on the impact of anonymized images on model training is presented by HUKKELAS [12]. The authors apply both realistic and non-realistic anonymization techniques. For traditional non-realistic methods, they use face and full-body blurring as well as mask-out. Realistic anonymization is implemented using DeepPrivacy2 for both face and full-body anonymization. Their results indicate that face anonymization has minimal to no impact on segmentation tasks for both realistic and traditional methods. However, for keypoint detection, traditional anonymization significantly degrades performance, whereas realistic anonymization results in a much smaller performance drop. When evaluating full-body anonymization, they observe a clear decline in performance for both traditional and realistic methods. However, realistic anonymization proves to be significantly better than traditional approaches.

While other works such as LEE and ZHOU [10, 11] suggest that larger models are less affected by anonymization, HUKKELAS [12] finds the opposite. Their analysis, is limited to ResNet and R-CNN architectures, leaving open the possibility that other architectures, such as YOLO as CNN, may behave differently.

Similar to LEE [10], they observe that anonymization impacts not only the altered classes but also unrelated ones. In full-body detection, they find no significant influence on large objects like buses, cars, motorcycles, trains, or trucks. However, performance decreases for smaller sized classes such as bicycles, which they theorize to be caused by overlapping of relevant regions. They highlight realistic anonymization as a "superior option", but emphasize that it is no substitute for real data, despite findings indicating that realistic anonymization can effectively replace original data in certain cases.

In their experiments, they do not anonymize the entire dataset. They exclude crowded scenes and images where objects are frequently in close relation to people (e.g., bicycles) to avoid relabeling the data, leading to the assumption that a

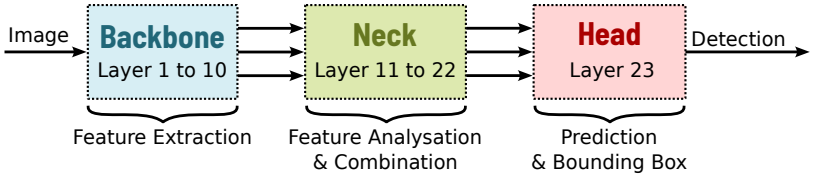

FIGURE 3. Particularly simplified scheme of YOLOv10, highlighting the task of specific layers with YOLOv10. For detailed architecture, see Appendix A.

mix of original and anonymized data is used for training if the whole training set is used.

2.3 ARCHITECTURE OF YOLOv10

To gain further insights into whether realistic anonymization fundamentally alters image features or merely requires fine-tuning of the classification step (see experiments in Section 3.6), a closer look at YOLO's architecture is necessary. The goal is to determine which parts are responsible for a certain task within the model's learning process. This section describes the architecture and tasks of individual layer groups in YOLOv10 according to [22] and [9].

YOLOv10's architecture can be divided into three key components, each with a distinct role in the detection process. A simplified representation of these responsibilities is shown in Figure 3. Due to the absence of a complete architecture diagram for YOLOv10 in the literature, our work provides an architectural overview, including individual layer types and corresponding layer numbers, in Appendix A.

Key Components

The backbone is responsible for extracting features from the input image. These detected features are passed to the neck and into a self-attention layer, which captures global-scale patterns.

Extracted features along with the global patterns provided by the self-attention layer are processed and combined within the neck. This enables a comprehensive feature combination before passing the data to the head.

The head determines the observed class types, defines bounding box areas, and outputs the final detections. It consists of two specialized components: a one-to-many head, which generates multiple predictions during training to improve model accuracy, and a one-to-one head, which selects the best-fitting detection during inference. This distributed approach eliminates the need for Non-Maximum Suppression (NMS), leading to improved latency compared to earlier YOLO versions — a key optimization goal of YOLOv10.

2.4 DATASETS

Publicly available datasets related to anonymization typically focus on the anonymization process itself. Which is evaluated by the level of achieved privacy [13]. Other datasets focus on specific computer vision tasks, but are commonly utilized to improve anonymization techniques. Currently, anonymization is not frequently used in datasets. There exist some, like NuScenes [24], A2D2 [25] or AViD [26] but within these, anonymization is restricted to the blurring of heads or license plates. However, there is a notable gap in specialized datasets designed to study the impact of different anonymization methods on computer vision tasks or AI training. Consequently, the most practical approach is to utilize a commonly used dataset and apply the desired anonymization method to it, as shown by LEE [10] and HUKKELAS [12].

This study, compiled a collection of relevant datasets applicable for various tasks and use cases, or enhancement of anonymization methods in Appendix B. Though, most of these datasets do not satisfy the requirements of our experiments, which are: the inclusion of persons performing diverse activities (preferably aligned with AAL or AD) in home or outdoor environments, annotations for everyday objects, and a manageable size allowing anonymization, training and evaluation within a reasonable timeframe. Among the options considered, MS COCO [27] is the only dataset meeting most of the criteria and is also employed in prior studies of LEE [10] and HUKKELAS [12].

2.5 CONTRIBUTIONS OF THIS WORK

This work aims to extend previous research of training and evaluating models on realistically full-body anonymized data. By combining and expanding evaluation schemes, it establishes a new systematic approach to assess the impact of anonymization on model training and performance, using object detection as a case study. This methodology is not limited to realistic anonymization, but is also applicable to other anonymization techniques or methods that alter specific image regions. The following section highlights the differences to prior works and summarizes our contributions.

Differences to Previous Works

In contrast to HUKKELAS [12], this work fully anonymizes the dataset for training without filtering the images for harder cases. While they exclude classes commonly associated with persons (e.g. bicycle or motorcycle), our approach enables a general evaluation taking all possible circumstances into account.

Additionally, it is also investigated if using the original annotations on anonymized data or relabeling influences detection performance. Whereas HUKKELAS [12] applies DeepPrivacy2 on different datasets (Cityscapes [23], BDD100K [28] and COCO) and filters problematic cases, this study employs the same anonymization framework but focuses on COCO, without removing problematic cases. Additionally,

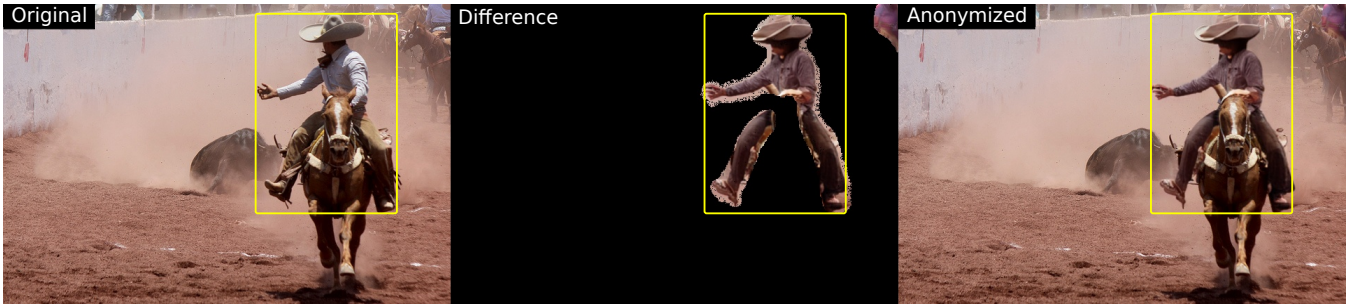

FIGURE 4. Difference images (middle) between original image from COCO (left) and anonymized with DeepPrivacy2 (right). The ground truth bounding box (yellow) shows no significant change in size or position for the generated person compared to the original image.

whereas ZHOU [11] uses ResNet/R-CNN architectures for segmentation, this work evaluates the performance of a CNN-based object detector (YOLO).

Further, while LEE [10] evaluates only on the original validation set, this study assesses model performance on both the original and anonymized validation sets. This comparison allows for an analysis of data exchangeability between different training and application data types.

In [10], LEE investigates which pixels of original and anonymized data contribute to classification. Findings indicate that models trained on original data rely on overall shape and contour for detections, whereas models trained on anonymized data interpret features and objects differently, though no further details are provided.

Our work aims to explore these effects by freezing layers of the model which are responsible for feature detection, combination and the final detection. Using this approach, it is expected to gain further insights whether realistic anonymization alters image features and therefore changes the models' detection process.

LEE [10] observes a correlation between decreased performance in non-anonymized objects and their frequent co-occurrence with the anonymized class. The present work aims to investigate this effect further by analyzing performance variations across different co-occurrence frequencies. Additionally, we hypothesize that object size plays a crucial role, as smaller objects are more likely to be altered or affected by anonymization. Therefore, this study examines not only the frequency of object appearances but also their size to better understand the impact of anonymization on detection performance.

Instead of using YOLOv8, as LEE [10], this work utilizes YOLOv10, which incorporates an updated head architecture for improved detection results.

To assess changes and imperfections introduced by anonymization, this work employs SSIM as an evaluation metric. The measure is also applied by ZHOU [11] to identify regions where anonymization is applied within a dataset. Their study, which focuses on the impact of anonymization on semantic segmentation, required detecting areas affected by anonymization. This provides a valuable assumption — generated content is not perfect and alters key image properties such as lighting and texture.

Building on this, our work adapts the use of SSIM from identifying anonymized regions to a method for quantifying changes introduced by anonymization, offering a broader perspective on its impact on model training and performance.

Overview of Contributions

- We establish a systematic evaluation approach to analyze anonymization methods, assessing both image modifications and their effects on model training and detection performance.
- For training and evaluation, the entire COCO dataset is anonymized without filtering out difficult cases, simulating real-world conditions.
- The anonymized COCO dataset is used to investigate training performance on anonymized data and test detection performance on the anonymized validation set.
- Further, we investigate the influence of model size on performance under realistic anonymization, contributing additional results to the contradiction between LEE, ZHOU [10, 11] and HUKKELAS [12].
- This study employs SSIM as a novel metric to quantify the changes introduced by anonymization.
- Building on LEE [10], our work explores the influence of frequent co-occurrence with the anonymized class 'person' in relation to object size.
- By identifying the most affected components of the detection pipeline under realistic anonymization we provide a direct extension to LEE [10], aiming to gain insights, that enhance model performance on anonymized data.
- Additionally, we investigate the exchangeability of original and anonymized data for training, offering insights for practical use cases.
- Finally, performance gains from relabeling anonymized data are demonstrated by comparing the use of the original ground truth with a ground truth adapted to anonymized data.

FIGURE 5. Difference images (middle) between original image from COCO (left) and anonymized with DeepPrivacy2 (right). Top: Changes occur not only for persons, but also in a small surrounding area, potentially changing properties of near objects.

3 EXPERIMENTS

This section presents experiments evaluating the impact of anonymization on model training and performance. It introduces the used data and describes the anonymized dataset's creation. The key investigations include: evaluating general detection performance and model size influence, assessing anonymization effects on an image level, comparing realistic to non-realistic anonymization, analyzing influence of object size and frequency, exploring fine-tuning with different freezing configurations, and assessing improvements through annotation correction for anonymized data.

3.1 DATASET

Due to the lack of datasets focusing on the influence of anonymization and the other reasons listed in Section 2.4, the MS COCO 2017 instance dataset is chosen as a base dataset for this work.

To generate the anonymized version of this dataset, the process visualized in Figure 1 is used. All images containing persons from both the training and validation subsets of COCO are retrieved and processed with DeepPrivacy2 in full-body anonymization mode.

As DeepPrivacy2 is an unsupervised approach to anonymization, a confirmation of its effectiveness is needed. This is verified by reviewing 100 randomly selected images, confirming that the original annotations maintain aligned within the anonymized images (compare Figure 4).

This process results in two datasets, each with training and validation data: COCO original and COCO full-body anonymized. For data generation, a system with an NVIDIA GeForce GTX 1070 (VRAM 8 GB) and CUDA 10.2 is used.

3.2 CHANGES OF ANONYMIZATION ON AN IMAGE LEVEL

The goal is to assess the changes beyond a count of pixels with differences. Since the anonymization process always affects an area surrounding the anonymized person (see Figure 5), a more advanced assessment method is required. A naive approach, such as computing the difference image between the original and anonymized versions of the same image or counting changed pixels, is considered to not provide sufficient insight. In particular, the grade of modifications must be observed, as neighboring textures or lighting are affected to varying degrees. As these changes are a potential explanation why objects frequently appearing alongside the 'person' class show a decline in detection performance, they need to be measured.

To address this, the Structural Similarity Index Measure (SSIM) [29] is employed, which differs from techniques like Mean Squared Error and other methods that estimate absolute differences. SSIM evaluates changes in luminance l, contrast c, and structure s between two images or neighborhoods x, y. The key idea is that pixels exhibit strong dependencies, especially within a neighborhood. These dependencies encode crucial structural information about objects in a scene. Therefore, properties such as the pixel mean μ_x, μ_y and variance σ_x, σ_y of x, y are incorporated into the computation. The full derivation can be found in [29], at this point a brief look at the main components of SSIM is provided.

The equations for l, c, and s are given by

$$l_{xy} = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\,, \tag{1a}$$

$$c_{xy} = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}\,, \tag{1b}$$

$$s_{xy} = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}\,, \text{where} \quad c_3 = \frac{c_2}{2}. \tag{1c}$$

The constants $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are used with the factors k_1, k_2 to weight the dynamic range L of pixel values, ensuring that l, c and s are balanced without overemphasizing any single component. The final SSIM score is derived as a weighted product of l, c, and s, resulting in

$$SSIM_{x,y} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}. \tag{2}$$

The values of SSIM are in the range of $[-1, 1]$. A value of 1 means perfect similarity, values of 0 indicate no similarity, negative values show an anti-similarity which can range up to -1 standing for absolute anti-correlation.

To evaluate SSIM, a closer look is taken at use case-specific classes for AAL and AD. Classes are divided into two groups for each use case:

- Change Group (CG): changes are acceptable
- No Change Group (NoCG): changes should be avoided

The resulting groupings of classes are shown in Table 1.

Using these groups, SSIM is calculated for the entire image containing the relevant classes. The basis of the calculation is the original COCO image and the anonymized version of the same image. Additionally, SSIM is calculated based on bounding boxes, comparing class relevant regions in the original image with the same areas in the anonymized version. This results in an SSIM score specific to each class relevant area. By comparing both methods, it is possible

Usecase \ Group	NoCG	CG
AD	car, stop sign, traffic light	cow, umbrella, bench
AAL	knife, bed, chair	potted plant, clock, tv

TABLE 1. Class groupings used to calculate SSIM scores for the use cases AD and AAL, divided into no change group (NoCG) and change group (CG)

YOLOv10 Size	Size (in million)
n	2.3
s	7.2
m	15.4
l	24.4
x	29.5

TABLE 2. Count of parameters for each of the used YOLOv10 model sizes.

Name	Description
Org	Model, Training data: Original COCO training set
Anon	Model, Training data: Full-Body anonymized COCO training set
on Org	Evaluation, Data: Original COCO validation set
on Anon	Evaluation, Data: Full-Body anonymized COCO validation set

TABLE 3. Naming convention used to define the type of training data and type of evaluation data used, e.g. a model named *Anon on Anon* is trained on full-body anonymized data and evaluated using the anonymized validation set.

Group	Riding	Equipment	Accessory
Class	bicycle motorcycle	skis snowboard baseball bat baseball glove skateboard surfboard tennis racket	backpack umbrella handbag tie suitcase

TABLE 4. Classes used by [10] to evaluate performance of trained models with different methods of anonymization.

to assess the changes related to persons and the impact of anonymization for each class.

3.3 GENERAL PERFORMANCE AND INFLUENCE OF MODEL SIZE

To investigate the impact of model size on the performance of anonymized models, YOLOv10 models of all sizes, from YOLOv10n to YOLOv10X, are trained using the Ultralytics toolbox [22]. The number of parameters is provided in Table 2.

Each model size is trained on both the original COCO training set and the full-body anonymized training set, resulting in two trained models per size, differing in training data. Both sets of models are evaluated on both the anonymized validation set and the original COCO validation set.

A naming convention is established for clarity, as given in Table 3. For example, a model named *Org on Org* is trained on the original COCO training set and evaluated on the original COCO validation set. This results in four possible model and evaluation combinations *Org on Org, Org on Anon, Anon on Org* and *Anon on Anon*. The *Org on Org* evaluation forms the base evaluation the other are compared to.

The training methodology is realized according to LEE [10]. They train the models from scratch using default parameters except for the optimization method, which is changed from 'auto' to 'SGD'. To retain comparability, we use the same configuration.

The only additional modification is an increased batch size of 40. To ensure reproducibility, the deterministic flag is set to 'True', and the training seed is fixed at 0. Evaluations are conducted across all training and evaluation data types, measuring mAP and AP over all classes and model sizes. Evaluations are realized using the COCO API [30]. Our models are trained on a system using an NVIDIA A40 (VRAM 48 GB), CUDA 12.1 and a total RAM of 512 GB.

3.4 REALISTIC ANONYMIZATION COMPARED TO NON-REALISTIC ANONYMIZATION

LEE [10] provides multiple tables with detailed mAP and AP values for the m-size YOLO model. These values are used to compare our trained models, which specialize in realistic full-body anonymization, with models trained using different non-realistic anonymization methods. Since most of their information concerns the m-size model, that size is the focus of further evaluations. Unfortunately, they supply detailed AP for a limited set of classes only, so the comparison remains restricted to the 14 classes listed in Table 4.

3.5 INFLUENCE OF OBJECT SIZE AND FREQUENCY

Assessment of the influence of object size and frequency are conducted on the four different model and evaluation data combinations. Following LEE [10], results are analyzed for models of size m.

To focus on the relationship between object frequency and performance impact, specific object classes are selected to represent different frequency levels. Classes frequently appearing with the 'person' class are categorized into three frequency groups based on the total number of images in which they appear together. These frequency distributions are illustrated by the blue bars in Figure 6. Borders of frequency ranges are chosen based on recognizable changes in classwise image count, resulting in definitions of:

- High frequency f_{high}: 9,000 down to 3,200 images
- Medium frequency f_{med}: 3,200 down to 1,400 images
- Low frequency f_{low}: less than 1,400 images

Since performance degradation is expected to depend not only on frequency but also on object size, the selected classes are further categorized by size, compared to humans:

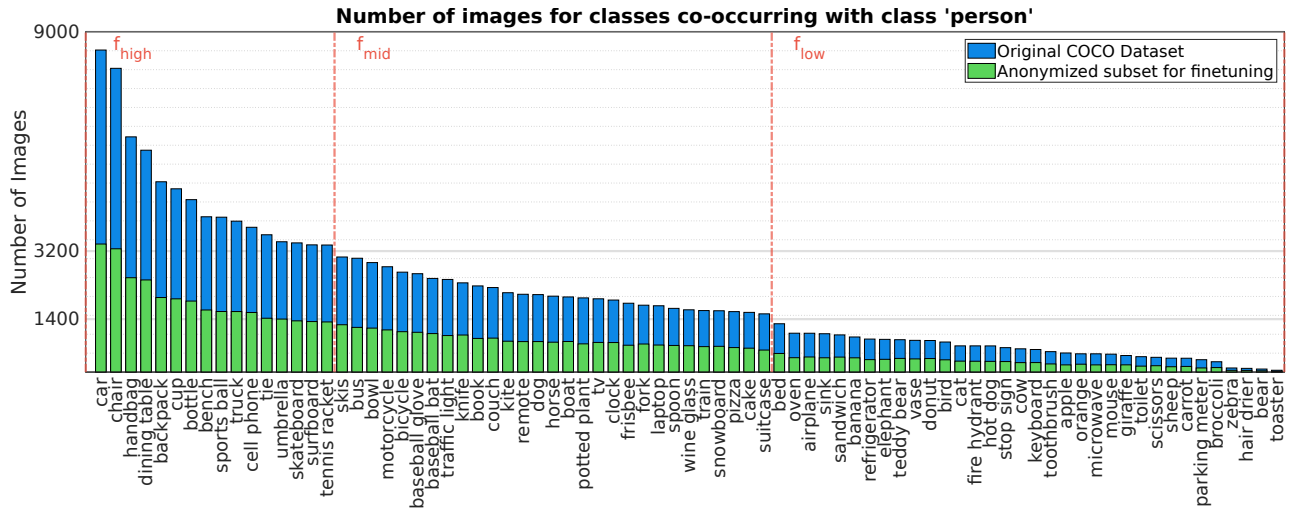

FIGURE 6. Distribution of images per class which is pictured with person class. Blue: original COCO train set, Green: used subset for fine-tuning with different configurations of frozen layers.

Size \ Frequency	low	medium	high
small	banana, vase, bird, toothbrush	book, bowl, baseball glove, clock	sports ball, cup, cell phone, bottle
medium	fire hydrant, microwave, toilet, keyboard	dog, laptop, baseball bat, suitcase	backpack, chair, umbrella, bench
large	refrigerator, elephant, bed, stop sign	couch, bus, horse, motorcycle	dining table, car, truck, surfboard

TABLE 5. Objects classified by size and frequency.

Definitions	Layers
BACKBONE	0 - 10
BACKBONE without PSA	0 - 9
NECK	11 - 22
NECK with HEAD	11 - 23
NECK with HEAD & PSA	10 - 23

TABLE 6. Definitions of frozen layer ranges. For architecture graph of YOLOv10 with layer (block) numbers, see Appendix A.

- Small objects: Items that can be easily carried in one hand or fit inside it.
- Medium-sized objects: Items reaching approximately knee or hip height when placed on the ground.
- Large objects: Items at least human-sized or significantly larger.

Using these classifications based on frequency and size, four classes per combination are selected. The chosen objects are listed in Table 5. To assess performance changes by anonymization, changes in AP are investigated for *Org on Anon, Anon on Org* and *Anon on Anon* relative to the *Org on Org* base evaluation.

3.6 INFLUENCE ON MODEL TRAINING

To determine where anonymization affects the model's learning process, the *Org* model of size m, is fine-tuned on anonymized data. To get an insight into which part of the model is influenced by the detection process, different configurations of frozen layers are trained.

By comparing the performance of different freezing strategies, it is possible to identify which step is mostly affected by anonymization and where its influence lies. Definitions of frozen layers can be seen in Table 6.

If fine-tuning of the neck (freezing the backbone) improves performance, it suggests that anonymization primarily affects analysis and combination of features rather than fundamentally altering the raw image. This would indicate that anonymized persons retain key natural characteristics, aligning with expectations based on the quality of anonymization, where essential features such as face, hair, arms, and legs seem clearly recognizable.

If fine-tuning of the backbone (freezing the neck) yields better results, it suggests that anonymization significantly alters the image at the pixel level, requiring the model to relearn feature extraction to improve performance. Additionally, the fine-tuned models are compared to the model trained entirely on anonymized data to assess whether pretraining on original data before fine-tuning on anonymized data provides an advantage when working with anonymized images.

Data Selection and Parameter for Fine-tuning

To prevent overfitting on already-learned classes which are probably unaffected by anonymization, a specific subset of the anonymized training data is selected. Only images containing anonymized persons and the selected classes are used. A random sample comprising 20 % of the full

FIGURE 7. Left: Examples of own dataset where anonymization is not introducing errors. Right: Examples of own dataset with problematic original ground truth. The top row shows the classes 'tie' (blue) and 'book' (pink) which are removed through anonymization. The bottom row visualizes how much the original 'person' bounding box (yellow) needs to be extended (yellow mask).

Parameter	Value
epochs	30
initial learning rate	0.001
final learning rate	0.01
momentum	0.85
weight decay	0.0007
warmup epochs	4
warmup momentum	0.8
optimizer	SGD
batch size	40

TABLE 7. Parameters used for fine-tuning. Settings for optimizer and batch size are adopted from the base model. Other values are based on suggestions from [31, 32] and auto-tune feature of Ultralytics. Parameters not listed here use default settings.

training set is drawn. To ensure class distributions are not over amplified, the image count per class in the fine-tuning subset is compared to the distribution in the full training set. Figure 6 shows this comparison, with the fine-tuning subset represented by green bars. Since the distributions align, the dataset is suitable for fine-tuning without introducing further biases. Used parameters for fine-tuning are listed in Table 7.

3.7 INFLUENCE OF LABEL ERROR

Using the original ground truth for anonymized data is not ideal, as errors may arise, particularly in cases where individuals are interacting with objects. During anonymization, there is no guarantee that these objects will be preserved (as seen in Figure 2), leading to labeling inconsistencies that can affect both training and evaluation.

In training, the model learns from altered pixel regions where objects have been removed or modified due to anonymization. These areas no longer resemble the same class of objects that remain unchanged, introducing inconsistencies. During evaluation, this issue presents two challenges: the model may correctly fail to detect an object that has been removed, yet this will be incorrectly counted as a false negative; alternatively, due to erroneous training data, the model might falsely detect objects which do not exist, leading to false positives.

These issues are briefly mentioned but not extensively explored by HUKKELAS [12], likely explaining why the authors evaluated their models on the original validation set rather than an anonymized one. Their decision to reduce anonymized frames to simpler cases with single individuals and exclude images with a high probability of person-object interactions (e.g. images with classes like bicycles and motorcycles) also aligns with this reasoning.

Due to the lack of specialized datasets for anonymization and AAL, the authors of the present work started to record an own custom dataset centered around AAL. Examples of the dataset are provided in Figure 7 (left). A small subset of 79 frames is used to assess the impact of relabeling. The original frames are annotated and relabeled after anonymization with DeepPrivacy2. To achieve this, the original ground truth is imported into the anonymized frames, compared, and corrected where necessary. Figure 7 (right) shows cases where these corrections need to be applied.

To quantify the effect of relabeling, models are evaluated on the anonymized subset using both the original ground truth and with corrected annotations. This comparison helps determine the extent to which annotation errors influence model performance and provides insights into the importance of accurate re-labeling in anonymized existing datasets.

Details of own Dataset

The dataset is recorded in an ambulatory care facility and includes 14 participants (3 male, 11 female). Each person takes various roles — caregiver, patient, visitor, or paramedic — and performs care-related tasks, such as measuring vital signs, wound dressing changes, movement therapy, fall incidents and assistance, as well as daily actions such as dialogs, reading, sleeping, eating and drinking. These actions are captured as continuous data across multiple scenes.

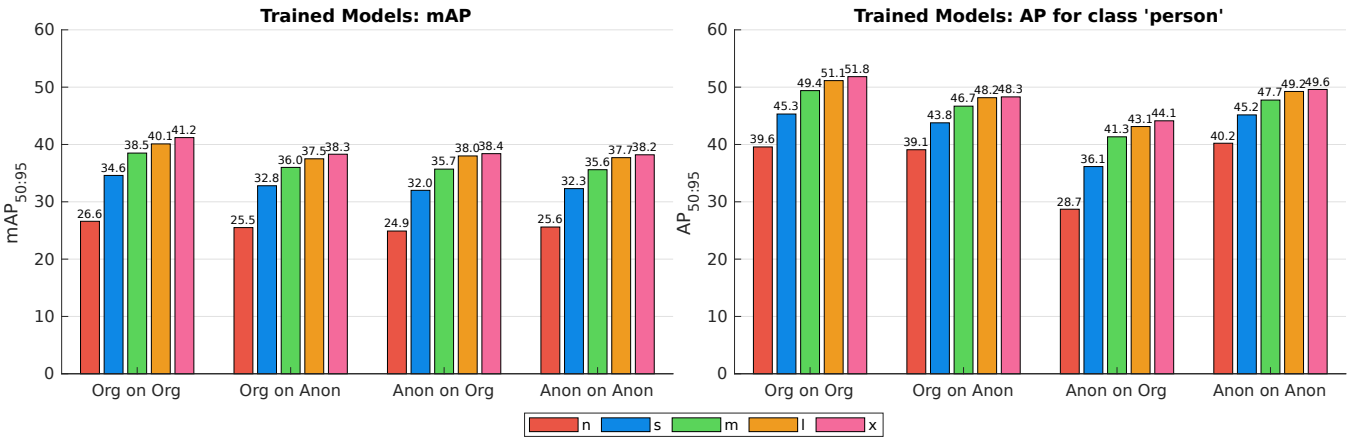

FIGURE 8. Left: Comparison of mAP of trained models on anonymized and original data across different model sizes. Right: AP of class 'person' for trained model types and sizes on anonymized and original data.

Model \ Size	n	s	m	l	x
Org on Org	26.6	34.6	38.5	40.1	41.2
Org on Anon	−1.1	−1.8	−2.5	−2.6	−2.9
Anon on Org	−1.7	−2.6	−2.8	−2.1	−2.8
Anon on Anon	−1.0	−2.3	−2.9	−2.4	−3.0

TABLE 8. Changes in mAP across model sizes for all training and evaluation types compared to the *Org on Org* base model. Larger models of all types show a larger performance decrease compared to smaller ones.

Model \ Size	n	s	m	l	x
Org on Org	26.6	34.6	38.5	40.1	41.2
Org on Anon	−1.1	−1.8	−2.5	−2.6	−2.9
Anon on Anon	25.6	32.3	35.6	37.7	38.2
Anon on Org	−0.7	−0.3	+0.1	+0.3	+0.2

TABLE 9. Changes in mAP across model sizes for exchanged evaluation data types compared to training type. The anonymized model *Anon* shows noticeably smaller changes than the model trained on original data (*Org*).

This dataset aims to provide a comprehensive basis for studying anonymization by capturing data across multiple modalities, offering valuable opportunities for future research. The sensor setup includes an Azure Kinect camera and a Blickfeld Cube1 LiDAR, with transformations between both sensors provided. Both sensors are mounted together, observing the scene from a top-corner viewpoint at 3 meters height with a 15° downwards angle, mimicking a real-world application setup.

Data is recorded using ROS2 [33], capturing the RGB image, depth image, and depth cloud of the Kinect, as well as the point cloud of the Cube1. Additionally, audio is recorded via the Kinect's microphone array. Recordings are stored as ROS2 bag and *.wav* files. For further processing, all ROS messages are extracted into individual image and *.pcd* files. The recorded scenes have a combined duration of about 30 minutes, with a total size of 1.2 TB.

The Kinect is configured to record at 15 FPS in 1440p resolution, which is later downsampled to 640×480 to align with the average COCO resolution.

LiDAR frames are recorded at 1.8 Hz using 150×150 scan lines at an angle spacing of 0.4° across a FOV of $70° \times 33°$ ($h \times v$).

At the current state, annotations are limited to 79 original and full-body anonymized RGB images and cover 18 objects: person, backpack, tie, umbrella, sports ball, bottle, cup, bowl, chair, dining table, bed, potted plant, clock, vase, laptop, remote, cell phone and book.

4 RESULTS

4.1 GENERAL PERFORMANCE AND INFLUENCE OF MODEL SIZE

To assess general performance and the influence of model size, the mAP of each model and evaluation combination (see Table 3) is compared across all trained model sizes. The anonymized class 'person' is examined to determine how performance changes through anonymization, offering insights on an important class for both defined use cases.

General Performance

Figure 8 (left) shows the mAP of all training and evaluation types across all classes. The mAP increases with model size for all model types. However, comparing absolute differences between model types reveals that larger models exhibit a bigger absolute decrease when compared to the base model *Org on Org*, as seen in Table 8. These results align with HUKKELAS [12] but contradict LEE [10] and ZHOU [11].

Compared to LEE [10], the differences remain small. Their work reports a decrease of about −25 across model sizes using distortion-based anonymization for keypoint detection, or a decrease ranging from −2.4 to −4.9 for object detection. The paper utilizes non-realistic anonymization, and evaluates on original images only. Achieving smaller reduction in performance by using realistic anonymization is the favorable option for anonymization, as outlined by HUKKELAS [12].

Nevertheless, performance drops with anonymized data

compared to *Org on Org*, whether used for evaluation or training, as seen in Table 8. For evaluations of models on their training data type, the anonymized model (*Anon on Anon*) shows a lower performance when compared to the base model (*Org on Org*). Other types show a slightly smaller decrease.

Comparing the performance of models on their own training type (e.g., *Org on Org*) with the opposite type (e.g., *Org on Anon*) in Table 9, reveals that the difference for models trained on anonymized data is small or even negligible. This indicates a certain exchangeability of data type for anon-trained models, though a small performance loss compared to their training data remains. Models trained on original data show a higher performance drop when switching data type.

This effect likely stems from imperfections introduced by anonymization. Embedding errors in training data is already known to make models more resilient (e.g., through data augmentation [34, 35]), which could explain this outcome. The extent of introduced imperfections and their possible effects are examined in Section 4.2.

Anonymized Class 'person'

For the anonymized class, similar observations as mentioned above apply and are visible in Figure 8 (right). However, some differences need to be highlighted.

For the 'person' class, there is a more clear difference in terms of data type exchangeability. For the original model *Org*, switching to anonymized data results in a performance loss (e.g., x-size *Org on Anon*: -3.5). Whereas, for an anonymized model *Anon*, switching to original data results in a greater loss (e.g., x-size *Anon on Org*: -5.5). In our use cases, these represent significant decreases for an important class, so switching between data types should be avoided within these use cases.

When maintaining the same training type within the evaluation (*Org on Org*, *Anon on Anon*), the difference between original and anonymized models becomes somewhat smaller (e.g., x-size: -2.2). By applying targeted tuning to the anonymized model and utilizing carefully crafted datasets, this small performance drop will potentially be reduced or even eliminated.

From an application point of view, in scenarios where anonymized data must be used, a model trained on anonymized data is preferable. In terms of person classification, the performance is notably better when both the training and inference are conducted on the same data type.

4.2 EFFECTS OF ANONYMIZATION ON AN IMAGE LEVEL

By comparing the SSIM score across the whole image and class specific bounding box areas, it is possible to assess the changes through anonymization which impact each of the defined classes (see list of classes in Table 1).

SSIM for whole Images

The achieved SSIM values for investigating changes across the entire image are visualized in Figure 9 (left). Although all used images contained the investigated class and persons, some classes reach a SSIM of 1, indicating no changes occurred through anonymization. One likely reason is that DeepPrivacy2 failed to anonymize persons, which does not necessarily imply poor performance. Images labeled containing "person" are not further verified for how much of the person is visible or how large the person is. COCO includes images containing only body parts or small persons, but still labeled with person annotations. These can remain undetected and are therefore not anonymized. An example of this issue is visualized in Appendix D.

The investigation of the AD classes reveals that all classes exhibit a mean SSIM greater than or equal to 0.9, indicating a high degree of similarity between the anonymized and original data. A significant proportion of the images achieve a SSIM greater than 0.8, further showcasing high similarity. Notably, classes where no changes are intended (AD NoCG) display high SSIM values with smaller variability, suggesting minimal alteration in the anonymized data. Even, classes where changes are acceptable (AD CG) show high SSIM values, though with greater variability.

Compared to the AD groups, all classes in the AAL group show a notably broader variance. For AAL, a larger portion of the data exhibits a lower SSIM, with values greater than 0.75. The mean SSIM within all classes fluctuates between 0.82 and 0.92. Additionally, the minimal SSIM values are significantly lower than those observed in the AD classes. These observations suggest that the anonymized images with selected AAL classes demonstrate less similarity to the original images compared to the AD classes, therefore undergo more significant changes through anonymization.

The lower SSIM values for the AAL group are likely attributed to the nature of the images in this dataset. AAL images have a higher probability to feature larger or closer individuals, as they are typically taken in home environments. As a result, people occupy a larger portion of the image, and the anonymization process has a more significant impact on these areas. In contrast, the AD group images are generally captured in outdoor environments, offering a wider field of view where individuals appear smaller, leading to less alteration by the anonymization process and, consequently, higher SSIM values.

SSIM for Bounding Boxes

To gain a clearer perspective on the impact of anonymization, it is essential to investigate the changes in the relevant areas corresponding to different classes. Using the same images, SSIM calculation is restricted to bounding boxes of single classes.

In Figure 9 (right), it is apparent that the classes 'car', 'stop sign', and 'traffic light' remain unchanged through anonymization, as reflected by their SSIM values of 1.0. This confirms that the NoCG group is not affected by anonymiza-

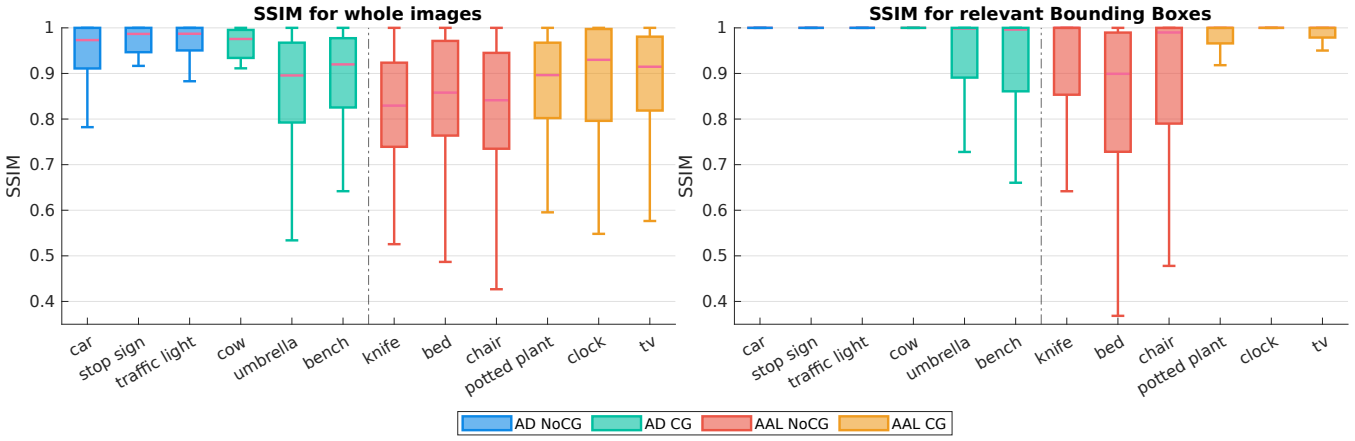

FIGURE 9. Results for SSIM calculation using the original image and the anonymized one. Left: SSIM for whole images which contain the relevant class. Right: SSIM-Results if only the respective ground truth bounding box area for calculation.

	car	stop sign	traffic light	cow	umbrella	bench	knife	bed	chair	potted plant	clock	tv
Org on Org	38.0	28.8	15.6	40.7	39.0	19.5	13.5	30.2	22.0	21.7	34.8	41.6
Org on Anon	0.0	+0.4	**−0.1**	**−0.6**	**−2.3**	+0.9	**−2.9**	**−1.3**	**−1.2**	+0.9	**−0.3**	**−0.8**

TABLE 10. Changes in AP (IoU = 50 to 95) for an YOLOv10 size m (original data trained) if evaluated on original data and anonymized data. Values of decreased AP are bold. Objects with decreased AP correspond to those with larger changes through anonymization, as seen in SSIM values in Figure 9.

tion. An explanation for this are the few instances of persons near these classes. Even within the AD CG group, the changes are minimal, with most data showing a similarity of approximately 0.9. This suggests that the use of anonymization in the AD use case does not significantly impact the use case relevant classes.

For AAL groups, classes in the NoCG category are noticeably more affected by anonymization, likely because they are more closely related to persons. This effect also appears for "umbrella" and "bench" in AD. Within AAL CG, values remain close to 1, even surpassing CG in AD.
These findings lead to the conclusion that anonymization for more human-centered applications like AAL is more problematic. Relevant classes show more pronounced changes and a larger decline in AP, which correlates with the bounding-box-based SSIM (see Table 10 and compare with Figure 9). These changes often stem from overlaps with generated content or partial contact of the object's outer edges, potentially altering important features. For instance, Figure 5 shows how anonymization removes the pointy edges of an umbrella.

As noted by LEE [10], models trained on original data often rely on shape and contour. Therefore, alterations can lower performance when using anonymized data. These observations explain effects noted in previous sections.

4.3 REALISTIC ANONYMIZATION COMPARED TO NON-REALISTIC ANONYMIZATION

To compare model performance regarding the anonymization method, models are compared with those of LEE [10] using the same set of 14 classes in Table 4. Figure 10 shows the AP for the evaluated classes at two IoU ranges: 1) IoU of 50 to 95 (left) and 2) IoU of 50 (right).

For an IoU of 50 to 95, our models show lower performance than the models from LEE [10]. The mean performance of the own models has already been discussed in Section 4.1, along with most differences between them. Notably, the *Org on Org* model exhibits the smallest variance in AP, while the variance of the reference models remains relatively consistent across all non-realistic anonymization methods.

When evaluation is performed at an IoU of 50, our models outperform those of LEE [10] (see right side of Figure 10). Better performance on an IoU of 50 indicates that they are more adept at detecting objects with less stringent overlap requirements. Conversely, performance dropping at higher IoU thresholds (50 to 95) suggests that the models have problems to accurately localize objects under stricter matching conditions, where predictions must closely align with ground truth annotations.

Lower IoU thresholds like 50 are more forgiving and suitable for scenarios where partial overlaps are common or where precise localization isn't critical. Higher IoU thresholds, such as 50 to 95, are stricter and better suited for tasks requiring precise object localization, such as in AD.

A possible explanation for the weaker performance is that LEE [10] uses non-realistic anonymization methods, which introduce abrupt changes in the anonymized region and disrupt natural pixel relationships. Such alterations may be easier for models to learn. These methods also tend to modify only the anonymized object.

In contrast, the realistic anonymization used in the present

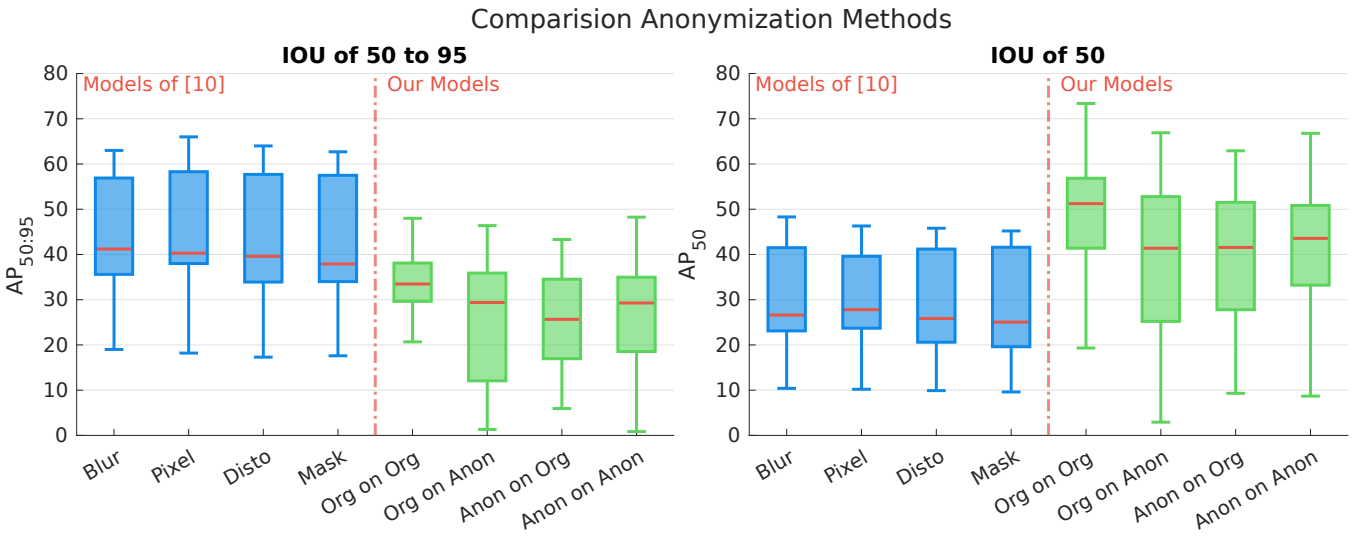

FIGURE 10. Comparison of models trained on anonymized data of [10] and our trained models. Both using YOLO size m. The study [10] uses original COCO data for evaluation. Our Models are trained on original data (Org) and anonymized (Anon) and evaluated using both the original validation set (on Org) and an anonymized one (on Anon). Left: Results for an IoU of 50 to 95. Right: Results for an IoU of 50. AP values are restricted to a set of 14 classes, listed in Table 4.

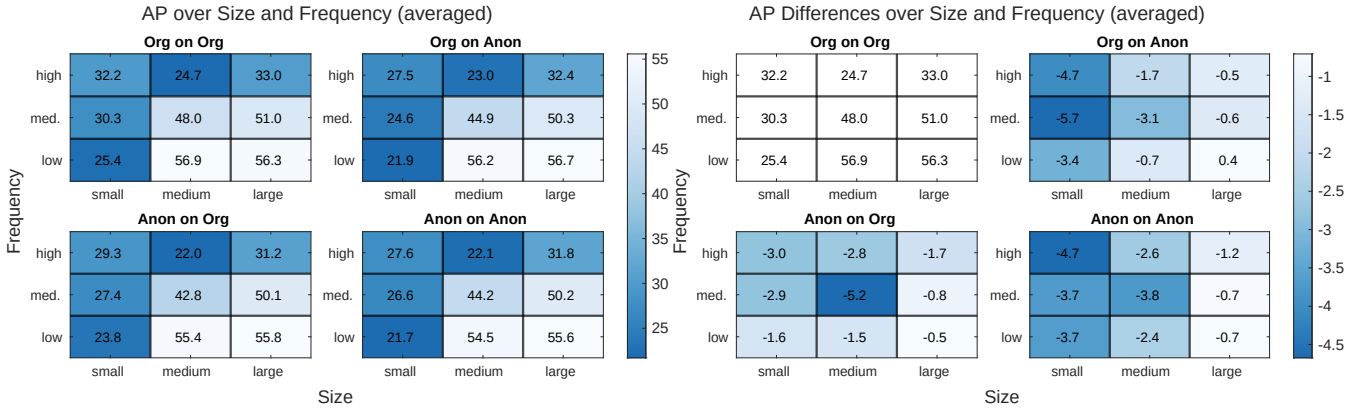

FIGURE 11. Left: AP of classes averaged for each Size-Frequency pair. Averages showing a similar behavior across all model and evaluation types. Absolute values of APs suggest a strong relation between lower AP and high frequencies as well for small objects. Right: Differences of averaged AP per Size-Frequency pair to base model *Org on Org*. Differences to the base model suggest a stronger relation between lower AP and size than between lower AP and frequency. Model size of pictured values: size m.

work, tries to generate natural-looking persons flowing seamless into the image. This alters a larger surrounding area, inducing errors and making them harder to detect. Another factor is the heavily restricted class set, which focuses on classes closely associated with humans — already shown to be problematic for the chosen anonymization method in Section 4.2.

However, these factors do not fully explain why even the base model (*Org on Org*) underperforms. A likely reason is that the training parameters and optimization strategies did not transfer as effectively to YOLOv10 as they did to YOLOv8, which was used by the comparison source.

4.4 INFLUENCE OF OBJECT SIZE AND FREQUENCY

Using the groupings from Table 5 defined in Section 3.5, the influence of object size and frequency with the anonymized class is investigated. Investigations are based on AP and changes in AP compared to *Org on Org*. The results are visualized in heatmaps. A lighter color indicates higher AP, where a darker color indicates lower AP. These values are averaged per size-frequency pairing. Individual class-AP are provided in Appendix C.

Figure 11 (left) shows the average AP for each size-frequency grouping. Objects with higher frequency to the 'person' class consistently have lower performance across all sizes, and smaller objects also exhibit generally lower performance. These findings hold for all model and evaluation types. This suggests they reflect a general performance trend, rather than changes related to anonymization.

To gain further insight into the relation of size and frequency to anonymization, Figure 11 (right) compares the change in AP by subtracting the *Org on Org* baseline from the

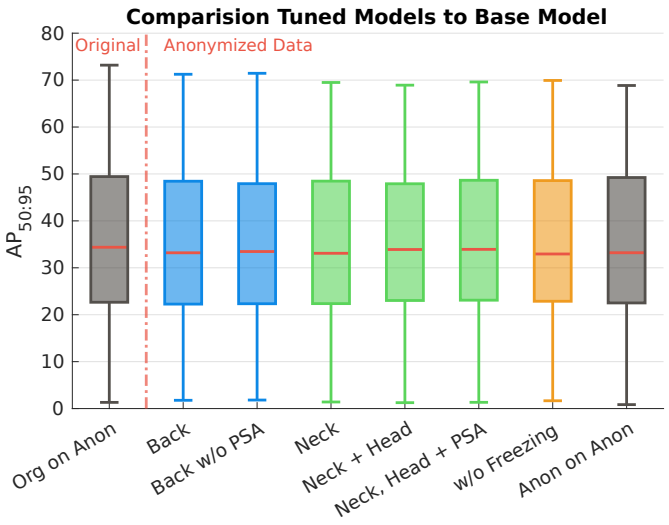

FIGURE 12. AP of all classes for different freezing strategies, compared to fine-tune base model (*Org* size m) and *Anon* model. All models are evaluated on anonymized data.

other three model-evaluation combinations. The coloration of tiles suggests a stronger connection between lower performance and smaller objects. However, higher frequency with the person class also contributes to reduced performance. Negative impacts are more pronounced for *Org on Anon* and *Anon on Anon* than for *Anon on Org*, suggesting that training on anonymized data provides increased resilience in cases of data type changes.

Like in LEE [10], classes with higher frequencies show worse performance than lower-frequency objects. Our data points to an even stronger link between low AP and smaller object sizes. The difficulty of smaller objects — already more challenging under normal conditions (compare left side of Figure 11, Heatmap of *Org on Org*) — is even more amplified under the influence of anonymization. They are more likely to be altered or removed through anonymization, especially if worn or handled by persons. This effect is evident when comparing objects fitting these conditions, e.g. the class 'tie' drops between *Org on Org* and *Org on Anon* from 31.4 AP to 1.3 AP.

4.5 INFLUENCE ON MODEL TRAINING

To determine where anonymization affects the model's learning, different freezing strategies for fine-tuning are compared. The AP comparison in Figure 12 shows that fine-tuning does not improve performance. Results with various frozen-layer configurations only converge toward the performance of the model trained on anonymized data, which is worse than the base model.

Among the strategies, no clear differences emerge: the mean and maximum AP values decrease for all variants, and there are only minor differences between models with a frozen backbone or neck, with the frozen neck models performing slightly worse.

Model	$AP_{50:95,\,person}$	$AP_{50,\,person}$
Org on Anon	46.7	65.8
BACKBONE	+1.1	+1.6
BACKBONE without PSA	+1.4	+2.4
NECK	+1.5	+2.5
NECK with HEAD	+1.5	+2.4
NECK with HEAD & PSA	+1.4	+2.4
No Freeze	**+1.7**	**+3.2**
Anon on Anon	+1.1	+1.7

TABLE 11. Differences in AP for the anonymized 'person' class across the different freezing strategies. Values are compared to the *Org* model, which is the base model used for fine-tuning. All models are evaluated on anonymized data. The best improvements are bold.

The model without any freezing is close in performance to the *Anon on Anon* model. Fine-tuning of an original model gives slightly better results than training directly on anonymized data, presumably due to the base model's prior knowledge. Though, differences are minimal.

The graph suggests a slight tendency for models with a frozen backbone to outperform those with a frozen neck, implying that realistic anonymization primarily affects feature analysis and combination. However, focusing on changes in the anonymized class points to a different, more relevant interpretation, since anonymization occurs there.

According to Table 11, improvements from fine-tuning on the anonymized 'person' class suggest that changes introduced by anonymization occur at the pixel level rather than the feature level. Fine-tuning the backbone leads to better results than freezing the backbone while fine-tuning the neck.

Not freezing at all delivers even better results, indicating that newly learned features also require different processing. These finding further align with LEE [10], where the authors note that anonymized models interpret features differently than models trained on original data. The initial assumption that generated humans retain key classification features (face, hair, arms, legs) does not seem to hold for classifying anonymized persons.

The general performance decrease on anonymized data, as observed in Figure 12, also appears in the other parts of this section. It highlights a wider issue when using an anonymized dataset with the original ground truth. Inserting generated content into the image sometimes partially or fully occludes other classes, leading to errors when combined with the original ground truth for evaluation purposes. Section 4.6 demonstrates the potential influence of correcting these errors.

4.6 INFLUENCE OF LABEL ERROR

A small sample from our in-development dataset is tested to see if relabeling is worthwhile. Both *Org* and *Anon* models are evaluated on the same sample. For evaluation, two ground-truth variants are used: one matching the original images and one with corrected annotations after anonymization.

Figure 13 shows the difference in AP. Both the AP for *Org* and *Anon* model either improve or remain unchanged,

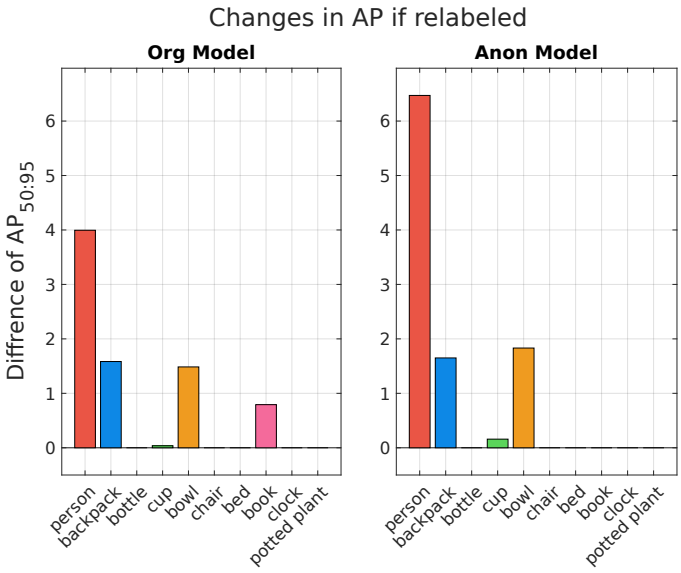

FIGURE 13. Changes of AP for classes where annotations are corrected for anonymized data. Graphs show differences of the same model evaluated using the ground truth of original data and ground truth with corrected annotations for anonymized data.

indicating that relabeling does not reduce AP for any of the evaluated classes. As expected, correcting labels on anonymized data leads to better results, particularly for small objects. There is also a measurable gain for the 'person' class. DeepPrivacy2 sometimes fails to correctly anonymize persons, e.g. when lying or sitting in bed, capturing parts of the bed or blanket in the generated instance (compare right side of Figure 7). If corrected to the new dimensions, these cases likely explain the improved performance for persons.

These findings suggest that correcting erroneous annotations yields a noticeable improvement in evaluation results for both Org and Anon training types. Therefore, models trained on anonymized data with corrected annotations will also produce better results, since the instances used for training fit the data supplied. The key result is that existing datasets cannot be directly used with anonymization without additional label adjustments.

5 CONCLUSION

Using the systematic approach portrayed in Section 3, this study identifies and investigates several problems that arise when working with anonymized data, especially regarding model training. These challenges include performance drops, difficulties in accurate object localization, and complications stemming from the nature of anonymized images.

Gaining these insights validates our systematic approach to analyzing anonymization methods successful. Nevertheless, issues like the labeling error and training parameters are identified. The main findings of this study as well as identified problems are summarized within this section.

Systematic approach

Training models of various sizes on original and anonymized data and evaluating both variants on anonymized and original data opens a wide range of evaluations and comparisons, offering the possibility to assess performance based on model size, as well as training and evaluation data types. This approach also provides insights into data exchangeability for real-world scenarios where only one model type (*Org* or *Anon*) is used. Employing SSIM as a metric assesses the changes introduced by anonymization and their influence on model performance. Freezing different configurations of layers during fine-tuning identifies the influence of anonymization on images and potential changes in the model's learning process. Investigating different object sizes and frequencies with the 'person' class highlights objects groups that are more heavily affected by anonymization.

The systematic approach allows us to answer the questions posed in Section 1, which renders our approach of evaluation successful. Applying it to further anonymization methods, networks, or computer vision tasks will yield valuable insights into the influence of anonymization.

Changes through Anonymization

Using SSIM as a metric to evaluate changes introduced by anonymization reveals significant differences between objects. For AD classes, it does not show relevant effects through realistic anonymization. In contrast, AAL applications, which are more human-centric, experience stronger changes to relevant classes and a greater performance decline. Overlaps frequently cause partial or complete occlusion by generated content, which is the primary source of problems.

The current state of realistic anonymization appears unsuitable for AAL environments, as the border area around generated persons also alters important features of surrounding objects. Additionally, DeepPrivacy2 fails to detect distant or partially occluded persons (e.g., a person lying in bed), making it ill-suited for both AAL and AD applications due to occlusion and distance issues.

General Performance and Model Size

With our systematic approach, it is possible to evaluate the performance of different model sizes and model types (*Org*, *Anon*) and their general performance on anonymized and original data.

Models trained on anonymized data perform better on them, than those trained on original data. Therefore, in scenarios where anonymized data is mandatory, models trained entirely on anonymized data are preferable. Regarding the anonymized class 'person', detection performs better when both training and evaluation data are of the same type.

In general, the trained models struggle to localize objects accurately, leading to lower performance under stricter matching conditions. In conclusion, the chosen realistic anonymization is not well suited for AD, as it requires a high degree of precision.

The differences in performance between models trained on realistic anonymized data and those trained on original data are generally small. However, evaluating on anonymized data or using models trained on anonymized data typically results in lower performance compared to models trained and evaluated on original data.

Compared to LEE [10], the models presented here perform worse, likely because the training parameters and optimization strategies did not transfer as effectively to YOLOv10 as they did to YOLOv8. Optimizing the training parameter for YOLOv10 will most likely yield clearer and improved results.

Larger models of all types exhibit a recognizable decrease in performance compared to smaller ones when working with anonymized data, rendering smaller models more robust to changes caused by anonymization. Therefore, the results comply with HUKKELAS [12], and contradict those of LEE [10] and ZHOU [11]. The contradiction with LEE [10] is especially peculiar, as they also use YOLO, unlike HUKKELAS [12] where ResNet/R-CNN architectures are used. There is currently no explanation for this. Further investigation is suggested.

Data Exchangeability

By switching the evaluation data to the opposite type of the training data (*Org on Anon*, *Anon on Org*), this study assesses data exchangeability.

The model trained on anonymized data shows noticeably smaller performance changes than the model trained on original data. We assume that anonymization-induced changes in classes strengthens the model trained on them, making it more robust. Nevertheless, the *Org* model performs better than the *Anon* model on both data types. Which indicates that in cases where both types need to be processed, either an original trained model should be used or multiple models, specialized to each type. For anonymized models on anonymized data, a small performance drop is to be expected compared to the original trained one.

Fine-tuning on Anonymized Data

Fine-tuning a model initially trained on original data offers the opportunity to compare its performance against a model trained directly on anonymized data.

We observed only a minor performance gain over the *Anon* model, likely due to knowledge already acquired by the base model. Although the original model still performs slightly better, the differences remain small. Parameter optimization for training and fine-tuning is likely to enhance the clarity of results.

Fine-tuned models only offer a minor improvement in comparison to the model directly trained on anonymized data. For the anonymized class 'person', the performance improves when the model is fine-tuned to handle anonymized data. Given these findings, fine-tuning is only worthwhile if the desired application is centered around humans.

Influence on the Model's Learning Process

Evaluating different freezing strategies provides insights into how the model processes anonymized images. Anonymization appears to affect pixel-level details rather than features. Freezing the neck (fine-tuning the backbone) performs better than freezing the backbone (fine-tuning the neck), indicating that changes occur primarily on the pixel level.

A detailed review of single layers (e.g., influence on the PSA layer) is not feasible, as differences between fine-tuned models remain minimal. Not freezing any part yields even better performance, suggesting newly learned features also need different processing, supporting the idea that anonymized models interpret data differently than those specialized in original data.

Influence of Object Size and Frequency

Classes with higher occurrence frequencies tend to perform worse than those with lower frequencies. This is also observed by LEE [10]. Our data indicates a stronger link between lower performance and smaller object sizes. Smaller objects are already challenging to detect, and anonymization amplifies this difficulty, particularly for items worn or held by a person, mainly because of overlapping or coverage by generation results.

Realistic Anonymization vs. Non-realistic Methods

A comparison between our results for realistic anonymization and those of LEE [10] for non-realistic methods reveals a significant performance difference. Based on different IOU ranges, our trained models show lower accuracy. We theorize that non-realistic anonymization introduces abrupt, unnatural alterations in the anonymized region, which seemingly simplifies learning for models.

In contrast, the realistic anonymization used here aims to create natural-looking persons blending into the scene, generating additional errors by modifying surrounding areas, potentially altering features or removing objects.

It is to highlight, that caution is warranted when comparing our results directly to those of LEE [10]. They solely evaluated training performance by training on anonymized data and evaluating on the unchanged coco validation set. In contrast, we additionally used the anonymized COCO validation set.

Observed differences between the studies motivated further exploration of the influence of anonymization on the validation set. Therefore, we investigated the correctness of the original ground truth on anonymized data using a subset of our dataset.

Label error

Experiments on our small data subset highlight that relying on the original ground truth is problematic when images are anonymized. Relabeling data for anonymized imagery is necessary to improve both training and evaluation.

Corrected annotations benefit smaller objects but also improve person-class performance. Using original ground

truth is believed to be one of the primary reasons for the anonymized model's lower performance. The issue is amplified by the lack of a dataset dedicated to anonymization, enforcing the use of datasets that are not designed for this purpose and require substantial work for relabeling and applying anonymization methods. This also leads to problems in comparing different studies.

Results across all experiments clearly highlight the influence of erroneous labels as the main source of problems for training or evaluation with anonymized data. Anonymization changes the shapes of objects or removes them from the image. Using original annotations leads to either erroneous training data or penalties for missing detections during the evaluation.

The only way to eliminate this issue is to re-label anonymized data, which likely improves training and evaluation results. These results amplify the need for specialized datasets for anonymization research, keeping these problems in mind and providing corrected annotations for anonymized data, like our sample dataset.

6 FUTURE WORK

All these challenges and findings summarized in Section 5 indicate a need for further research focusing on three priorities: improving realistic anonymization methods, adapting models to anonymized data, and providing datasets suitable for anonymization research to advance both of the first two goals.

Improving Realistic Anonymization

In DeepPrivacy2, minimizing the bordering region around the anonymized person can reduce alterations to other classes, and will therefore enhance subsequent processing.

Since DeepPrivacy2 aims for a seamless integration of anonymized regions into the surrounding image, further enhancements (e.g. adapting to lighting and other image factors) would be beneficial. As occlusion remains a major challenge for anonymization, this study highlights the importance of preserving objects within anonymized areas and calls for research featuring this problem.

Additionally, as realism is potentially important for further processing after anonymization, DeepPrivay2's tendency to context related generation is a feature likely needed. Refining this characteristic in future iterations provides the capability of context-aware identity generation.

However, even full-body anonymization does not inherently guarantee privacy as other identifiable features, like gait, may remain. Exploring the extent to which DeepPrivacy2 alters gait, alongside its prompt generation and sequence anonymization features, could yield valuable improvements.

Adapting Models to Anonymized Data

As models trained on anonymized data classify differently than those trained on original data, a deeper exploration of how these models make decisions could help tune them for anonymized inputs. Findings may guide the creation of specialized datasets that emphasize the distinct detection processes in anonymized scenarios. Even adjustments to model architecture may further improve performance.

Providing Datasets suitable for Anonymization Research

There is a notable gap in computer vision datasets designed specifically with anonymization in mind. Such a dataset would require rich annotations (objects, segmentation masks, poses, human actions, emotions, context) while also being use case–specific, given that task performance can strongly depend on objects near anonymized classes.

Development of this resource is challenging not only because of the considerable time and effort required, but also due to decisions on which anonymization techniques to employ. Correcting annotations for each chosen method adds further complexity.

Advancements in these priorities boost the research and introduction of privacy and law respecting AAL and AD applications. This work sees a need for privacy-by-design systems, which include anonymization. These are needed to produce data usable for follow-up algorithms. Given the findings of this and prior works, further advancements in the field of anonymization are important to establish life-changing technologies like AAL and AD.

REFERENCES

[1] A. Sweeting, K. A. Warncken, and M. Patel, "The Role of Assistive Technology in Enabling Older Adults to Achieve Independent Living: Past and Future," *Journal of Medical Internet Research*, vol. 26, e58846, Jul. 30, 2024, ISSN: 1438-8871. DOI: 10.2196/58846. [Online]. Available: https://www.jmir.org/2024/1/e58846 (visited on 01/22/2025).

[2] E. Schoitsch, "Autonomous Vehicles and Automated Driving: Status, Perspectives, and Societal Impact," in *IDIMT-2016: Information Technology, Society and Economy Strategic Cross-Influences: 24th Interdisciplinary Information Management Talks, Sept. 7-9, 2016, Poděbrady, Czech Republic*, ser. Schriftenreihe Informatik 45, P. Doucek, G. Chroust, and V. Oškrdal, Eds., Linz: Trauner Verlag, 2016, pp. 405–423, ISBN: 978-3-99033-869-8.

[3] S. Stecklow, W. Cunningham, H. Jin, and S. Stecklow, "Tesla workers shared sensitive images recorded by customer cars," *ReutersTechnology*, Apr. 6, 2023. [Online]. Available: https://www.reuters.com/technology/tesla-workers-shared-sensitive-images-recorded-by-customer-cars-2023-04-06/ (visited on 02/01/2025).

[4] K. Hill, "Your Car Is Tracking You. Abusive Partners May Be, Too.," *The New York TimesTechnology*, Dec. 31, 2023, ISSN: 0362-4331. [Online]. Available:

https://www.nytimes.com/2023/12/31/technology/car-trackers-gps-abuse.html (visited on 02/01/2025).

[5] N. Bookert, M. Almousa, and M. Anwar. "Inclusive Privacy Design for Older Adults Living in Ambient Assisted Living." arXiv: 2207.09592 [cs]. (Jul. 19, 2022), [Online]. Available: http://arxiv.org/abs/2207.09592 (visited on 01/06/2025), pre-published.

[6] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council. of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, OJ L 119, 4.5.2016, p. 1–88, May 4, 2016. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj (visited on 04/13/2023).

[7] K. Zhang *et al.*, "Deblurring by Realistic Blurring," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 2734–2743, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.00281. [Online]. Available: https://ieeexplore.ieee.org/document/9156306/ (visited on 01/28/2025).

[8] H. Hukkelas and F. Lindseth, "DeepPrivacy2: Towards Realistic Full-Body Anonymization," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 1329–1338, ISBN: 978-1-6654-9346-8. DOI: 10.1109/WACV56688.2023.00138. [Online]. Available: https://ieeexplore.ieee.org/document/10030153/ (visited on 12/25/2024).

[9] A. Wang *et al.* "YOLOv10: Real-Time End-to-End Object Detection." arXiv: 2405.14458 [cs]. (Oct. 30, 2024), [Online]. Available: http://arxiv.org/abs/2405.14458 (visited on 02/03/2025), pre-published.

[10] J. H. Lee and S. J. You, "Balancing Privacy and Accuracy: Exploring the Impact of Data Anonymization on Deep Learning Models in Computer Vision," *IEEE Access*, vol. 12, pp. 8346–8358, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3352146. [Online]. Available: https://ieeexplore.ieee.org/document/10387326/ (visited on 01/05/2025).

[11] J. Zhou and J. Beyerer, "Impacts of Data Anonymization on Semantic Segmentation," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, Aachen, Germany: IEEE, Jun. 5, 2022, pp. 997–1004, ISBN: 978-1-6654-8821-1. DOI: 10.1109/IV51971.2022.9827262. [Online]. Available: https://ieeexplore.ieee.org/document/9827262/ (visited on 01/05/2025).

[12] H. Hukkelås and F. Lindseth. "Does Image Anonymization Impact Computer Vision Training?" arXiv: 2306.05135 [cs]. (Jun. 8, 2023), [Online]. Available: http://arxiv.org/abs/2306.05135 (visited on 01/05/2025), pre-published.

[13] S. Ravi, P. Climent-Pérez, and F. Florez-Revuelta, "A review on visual privacy preservation techniques for active and assisted living," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14 715–14 755, Jul. 4, 2023, ISSN: 1573-7721. DOI: 10.1007/s11042-023-15775-2. [Online]. Available: https://link.springer.com/10.1007/s11042-023-15775-2 (visited on 04/02/2024).

[14] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 8877–8886, ISBN: 978-1-7281-4803-8. DOI: 10.1109/ICCV.2019.00897. [Online]. Available: https://ieeexplore.ieee.org/document/9008540/ (visited on 01/28/2025).

[15] D. Rozumnyi, M. R. Oswald, V. Ferrari, J. Matas, and M. Pollefeys, "DeFMO: Deblurring and Shape Recovery of Fast Moving Objects," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA: IEEE, Jun. 2021, pp. 3455–3464, ISBN: 978-1-6654-4509-2. DOI: 10.1109/CVPR46437.2021.00346. [Online]. Available: https://ieeexplore.ieee.org/document/9577579/ (visited on 01/28/2025).

[16] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, "I Know That Person: Generative Full Body and Face De-identification of People in Images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1319–1328, ISBN: 978-1-5386-0733-6. DOI: 10.1109/CVPRW.2017.173. [Online]. Available: http://ieeexplore.ieee.org/document/8014907/ (visited on 04/02/2024).

[17] H. Hukkelås, M. Smebye, R. Mester, and F. Lindseth. "Realistic Full-Body Anonymization with Surface-Guided GANs." arXiv: 2201.02193 [cs]. (Jun. 1, 2023), [Online]. Available: http://arxiv.org/abs/2201.02193 (visited on 01/22/2025), pre-published.

[18] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks," in *2020 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, pp. 5446–5455. DOI: 10.1109/CVPR42600.2020.00549. arXiv: 2005.09544 [cs]. [Online]. Available: http://arxiv.org/abs/2005.09544 (visited on 01/22/2025).

[19] J. Li *et al.*, "DSFD: Dual Shot Face Detector," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 5055–5064, ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00520. [Online]. Available: https://ieeexplore.ieee.org/document/8954268/ (visited on 01/22/2025).

[20] N. Neverova, D. Novotny, V. Khalidov, M. Szafraniec, P. Labatut, and A. Vedaldi. "Continuous Surface Embeddings." arXiv: 2011.12438 [cs]. (Nov. 24, 2020), [Online]. Available: http://arxiv.org/abs/2011.12438 (visited on 01/22/2025), pre-published.

[21] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2980–2988, ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.322. [Online]. Available: http://ieeexplore.ieee.org/document/8237584/ (visited on 01/22/2025).

[22] G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*, version 8.0.0, Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[23] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] H. Caesar *et al.*, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 11 618–11 628, ISBN: 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.01164. [Online]. Available: https://ieeexplore.ieee.org/document/9156412/ (visited on 02/03/2025).

[25] J. Geyer *et al.*, "A2D2: Audi autonomous driving dataset," 2020. arXiv: 2004.06320 [cs.CV]. [Online]. Available: https://www.a2d2.audi.

[26] A. J. Piergiovanni and M. S. Ryoo. "AViD Dataset: Anonymized Videos from Diverse Countries." arXiv: 2007.05515 [cs]. (Nov. 3, 2020), [Online]. Available: http://arxiv.org/abs/2007.05515 (visited on 02/03/2025), pre-published.

[27] T.-Y. Lin *et al.* "Microsoft COCO: Common Objects in Context." arXiv: 1405.0312 [cs]. (Feb. 21, 2015), [Online]. Available: http://arxiv.org/abs/1405.0312 (visited on 01/27/2025), pre-published.

[28] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[29] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, ISSN: 1057-7149. DOI: 10.1109/TIP.2003.819861. [Online]. Available: http://ieeexplore.ieee.org/document/1284395/ (visited on 01/29/2025).

[30] *Cocodataset/cocoapi*, cocodataset, Jan. 30, 2025. [Online]. Available: https://github.com/cocodataset/cocoapi (visited on 02/01/2025).

[31] Y. Bengio. "Practical recommendations for gradient-based training of deep architectures." arXiv: 1206.5533 [cs]. (Sep. 16, 2012), [Online]. Available: http://arxiv.org/abs/1206.5533 (visited on 02/17/2025), pre-published.

[32] "Momentum," in *Deep Learning* (Adaptive Computation and Machine Learning), Adaptive Computation and Machine Learning. Cambridge, Mass: The MIT press, 2016, pp. 292–296, ISBN: 978-0-262-03561-3.

[33] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot operating system 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, eabm6074, 2022. DOI: 10.1126/scirobotics.abm6074. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abm6074.

[34] M. Liu, W. Hong, W. Pan, and C. Feng, "A Robustness-Oriented Data Augmentation Method for DNN," in *2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, Hainan, China: IEEE, Dec. 2021, pp. 1–8, ISBN: 978-1-6654-7836-6. DOI: 10.1109/QRS-C55045.2021.00011. [Online]. Available: https://ieeexplore.ieee.org/document/9741895/ (visited on 02/08/2025).

[35] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. "Data Augmentation Can Improve Robustness." arXiv: 2111.05328 [cs]. (Nov. 9, 2021), [Online]. Available: http://arxiv.org/abs/2111.05328 (visited on 02/08/2025), pre-published.

[36] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A Dataset for Large Vocabulary Instance Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 5351–5359, ISBN: 978-1-7281-3293-8. DOI: 10.1109/CVPR.2019.00550. [Online]. Available:

https://ieeexplore.ieee.org/document/8954457/ (visited on 01/27/2025).

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255, ISBN: 978-1-4244-3992-8. DOI: 10.1109/CVPR.2009.5206848. [Online]. Available: https://ieeexplore.ieee.org/document/5206848/ (visited on 01/27/2025).

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[39] A. Kuznetsova *et al.*, "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.

[40] H. Hukkelås, R. Mester, and F. Lindseth, "DeepPrivacy: A Generative Adversarial Network for Face Anonymization," in *Advances in Visual Computing*, G. Bebis *et al.*, Eds., vol. 11844, Cham: Springer International Publishing, 2019, pp. 565–578, ISBN: 978-3-030-33719-3 978-3-030-33720-9. DOI: 10.1007/978-3-030-33720-9_44. [Online]. Available: http://link.springer.com/10.1007/978-3-030-33720-9_44 (visited on 12/25/2024).

[41] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015.

[42] S. Abu-El-Haija *et al.* "YouTube-8M: A Large-Scale Video Classification Benchmark." arXiv: 1609.08675 [cs]. (Sep. 27, 2016), [Online]. Available: http://arxiv.org/abs/1609.08675 (visited on 01/27/2025), pre-published.

[43] K.-S. Wong, N. A. Tu, A. Maratkhan, and M. Demirci, "A Privacy-Preserving Framework for Surveillance Systems," in *2020 the 10th International Conference on Communication and Network Security*, Tokyo Japan: ACM, Nov. 27, 2020, pp. 91–98, ISBN: 978-1-4503-8903-7. DOI: 10.1145/3442520.3442524. [Online]. Available: https://dl.acm.org/doi/10.1145/3442520.3442524 (visited on 01/27/2025).

[44] X. Liang, L. Lin, W. Yang, P. Luo, J. Huang, and S. Yan, "Clothes Co-Parsing Via Joint Image Segmentation and Labeling With Application to Clothing Retrieval," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1175–1186, Jun. 2016, ISSN: 1520-9210, 1941-0077. DOI: 10.1109/TMM.2016.2542983. [Online]. Available: https://ieeexplore.ieee.org/document/7434660/ (visited on 01/27/2025).

[45] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[46] L. Xie, *Hardhat*, application/x-rar-compressed, version 1.0, Harvard Dataverse, 2019. DOI: 10.7910/DVN/7CBGOS. [Online]. Available: https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/7CBGOS (visited on 01/27/2025).

.

APPENDIX A YOLOv10 ARCHITECTURE GRAPH

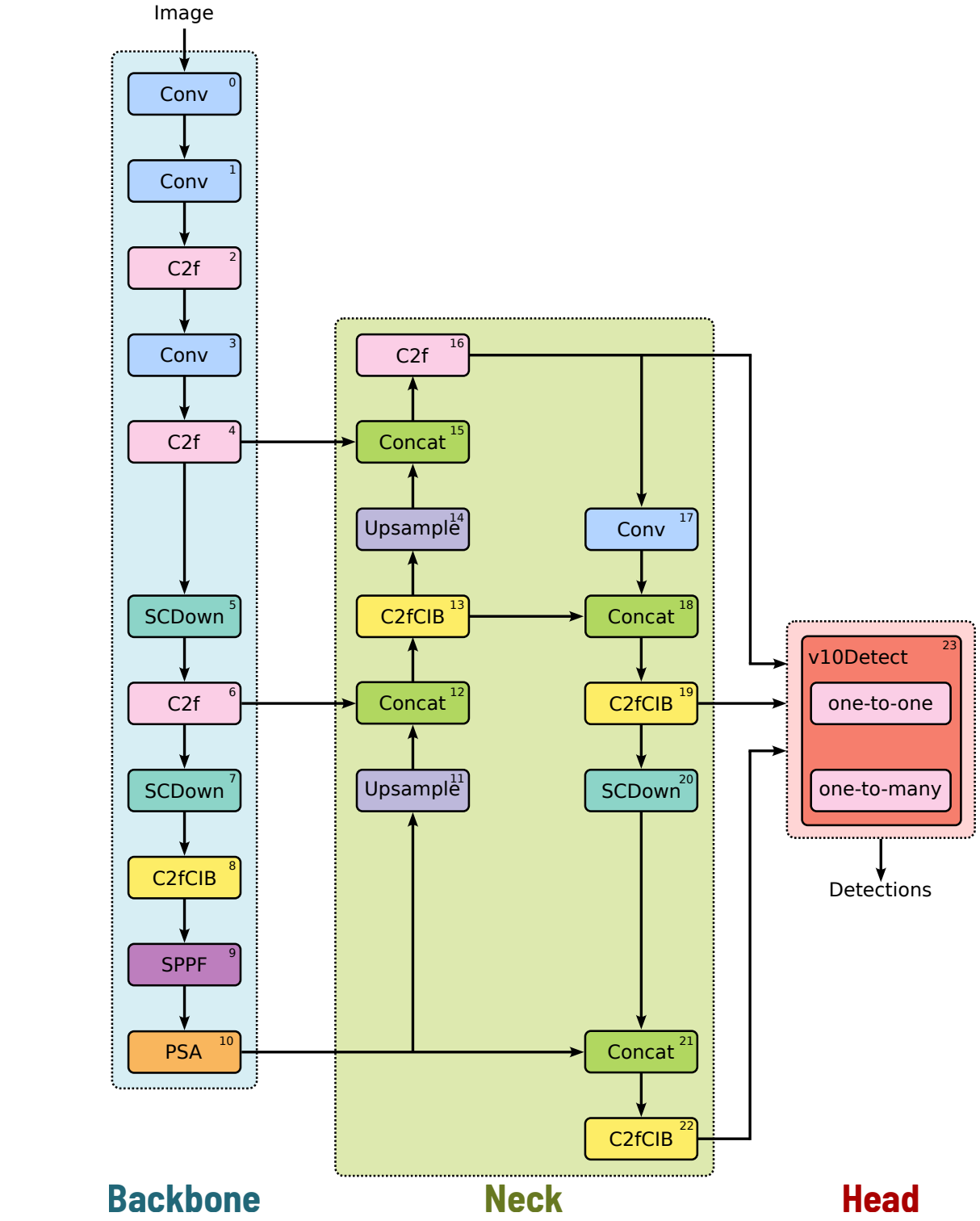

FIGURE 14. Architecture of YOLOv10 size m, based on layer configuration file from [22] and YOLOv10 paper [9]. Layers types and number according to [22]. Please notice that architecture has different types for some layers for other model sizes.

APPENDIX B DATASETS

Datasets for object detection and segmentation tasks

MS COCO [27]: A widely used dataset focusing on object detection, segmentation, and captioning. It features scenes from everyday life, including homes, urban and natural environments, various human activities, and crowded settings. The dataset comprises 330,000 images, with over 200,000 labeled images covering 80 object categories (91 if 'stuff' categories are included), and 250,000 people annotated with keypoints. It includes object segmentation, contextual information, and 5 captions per image.

LVIS (Long-tailed Visual Instance Segmentation) [36]: A large-scale dataset for instance segmentation, featuring 2 million masks across 1.2 million images and over 1,200 object categories. LVIS provides rich contextual information, long-tail distribution of object frequencies, and detailed annotations, including object attributes and multi-labels per object. It features scenes from diverse environments such as urban, indoor, and natural settings.

ImageNet [37]: ImageNet is a large-scale dataset designed for visual recognition tasks, featuring over 14 million images labeled across more than 20,000 categories. It emphasizes object classification, detection, and localization, with its most widely used subset consisting of 1,000 classes spanning approximately 1.43 million images. Images are richly annotated, primarily depicting single objects in varied contexts rather than complex scenes with multiple labeled objects or people.

Pascal VOC 2012 Dataset [38]: The Pascal VOC 2012 dataset is a benchmark for object detection, segmentation, and action classification tasks. It consists of >11,000 images with over 27,000 annotated objects across 20 categories, including animals, vehicles, and household items. Annotations include bounding boxes, pixel-wise segmentation, and object class labels. Images often depict realistic settings, such as streets, parks, indoor environments, or natural landscapes, and frequently include multiple objects in varied poses. The dataset supports multi-label and complex scene analysis.

Open Images [39]: A large-scale dataset from Google featuring approximately 9 million images with diverse annotations. These include object bounding boxes, segmentation masks, visual relationships and properties (e.g., 'woman playing guitar', 'beer on table'), localized narratives (detailed content descriptions of foreground, middle, and background), and point-level labels (specific pixel-level annotations). The number of classes and total image counts vary significantly by annotation type — for instance, 350 classes for instance segmentation, 600 classes for bounding boxes, and 1,466 different descriptions for relationships. The dataset spans a wide range of content, from single objects to complex scenes across various topics.

Datasets to enhance realistic anonymization results

Realistic anonymization is a complex task that requires expertise in pose-related generation, clothing synthesis, human action modeling, and more. Consequently, the development of these methods often relies on a diverse selection of datasets. The following section provides an overview of datasets, may be valuable for anonymization or enhancing anonymization techniques.

Flickr Diverse Humans Dataset [40]: This dataset consists of real-life images featuring centered single humans, annotated with dense pose (CSE), keypoints, and segmentation masks. It lacks additional annotations, such as object segmentation, focusing solely on human-centric data. With 1.5 million images for training and 30,000 for validation, it serves as a valuable resource for improving the synthesis and analysis of human poses. The dataset contributed to the development of the first version of DeepPrivacy [40] and was released with it.

CelebFaces Attributes [41]: CelebFaces Attributes features facial attributes annotated with approximately 40 flags, such as gender, age, and expressions. It contains 200,000 images and is particularly interesting for tasks involving facial editing, and synthesis of facial features.

YouTube 8M video dataset [42]: Comprising around 500,000 hours of video streams annotated with 4,800 labels for video categorization, this dataset is valuable for studying context preservation. While it provides extensive content, labeling individuals within the videos poses a challenge. The dataset has been utilized in privacy preservation research, such as in [43], with a focus on facial anonymization.

Clothing Co-Parsing dataset [44]: This dataset comprises 2,000 high-resolution street fashion images annotated with 59 labels for various clothing types, along with 1,000 images featuring pixel-level annotations. It provides detailed segmentations of individual garments and skin, making it a valuable resource for improving clothing generation techniques. The dataset has already been applied in a realistic anonymization method, as demonstrated in [16], to enhance the synthesis of clothing in anonymized images.

Human3.6M dataset [45]: This dataset provides extensive and diverse 3D human pose data, including 3.6 million annotated 3D poses paired with corresponding images. Poses are captured by 11 professional actors performing 17 distinct scenarios, such as eating, making purchases, pointing in directions, and phone conversations. Captured poses are used to generate human 3D models into real scenery. Annotations also include background subtraction, bounding boxes for individuals, and detailed pose information. As well as the Clothing Co-Parsing dataset, it has been applied in [16].

Hardhat dataset [46]: This dataset focuses on hardhat detection in industrial environments, contributing to workplace privacy and safety-related object detection tasks. It includes approximately 12,000 images with bounding box annotations specifically for hardhats, providing a resource for applications in industrial safety monitoring.

APPENDIX C DETAILED RESULTS

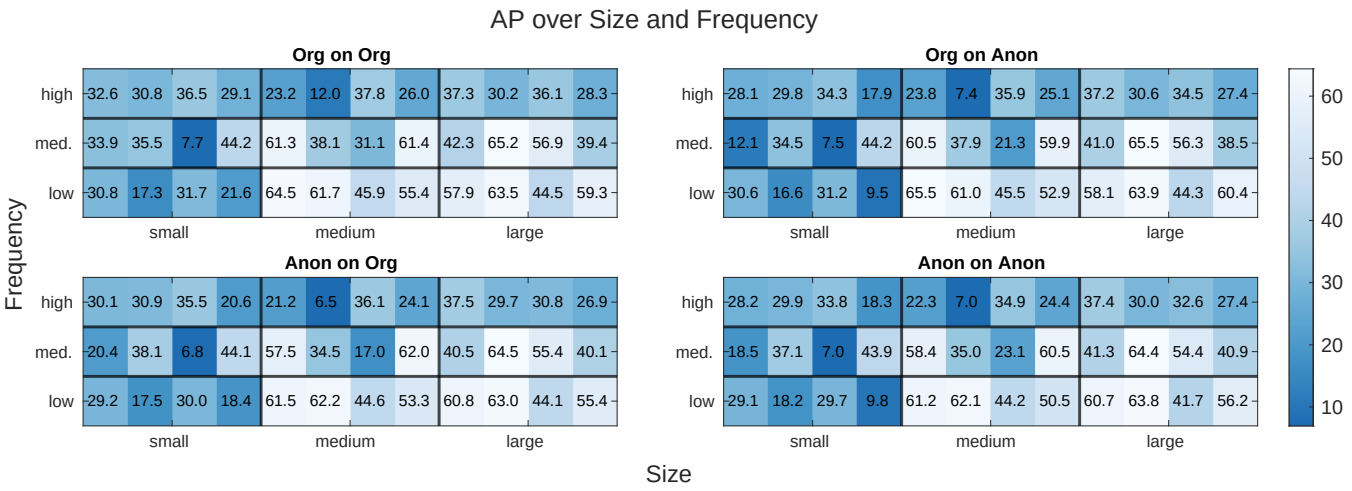

FIGURE 15. AP of single Objects within Size-Frequency pairs.

APPENDIX D ADDITIONAL IMAGES

FIGURE 16. Example, where DeepPrivacy2 is unable to anonymize persons as they are too small to be found by its detectors. Nevertheless, the persons are labeled within the COCO annotations. Even for the human eye, they are hard to spot.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit mit dem Titel

Computer Vision with Anonymized Data:
A Systematic Approach for Evaluation using Realistic Anonymization

selbstständig und ohne unerlaubte fremde Hilfe angefertigt, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet und die den verwendeten Quellen und Hilfsmitteln wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht oder die Arbeit in ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt habe.

Weingarten, 26.03.2025

Ort, Datum

Unterschrift