Privacy-Aware Perception System for Robotics

Baek Cheol Kim u7617018
Chenyang Li u7631563
Chen Yao u7722352
Qiance Zhou u7520051

College of Engineering, Computing & Cybernetics, The Australian National University

1 Introduction

1.1 Background

Robotics is playing a growing role in our daily lives, with robots now handling tasks that span from industrial production to home assistance. Integrating robots into public and private environments has introduced new challenges related to privacy and security (Kirschgens et al., 2021), as robots perception system may capture and process privacy-sensitive information such as human facial data throughout their tasks shown in Figure 1. Facial scanning, even when used incidentally for navigation, can violate privacy, raise ethical issues, and increase security risks. (Girasa, 2020).

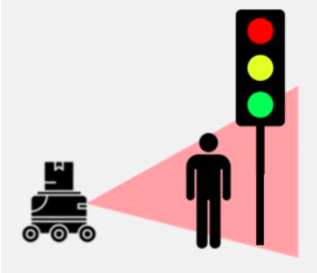


Figure 1: A robot navigating a public space with its perception system focused on a pedestrian. The robot's field of view is highlighted by the pink triangle, capturing the visual image of the individual including facial details. This setup emphasises the robot's potential to collect information within its field of view, raising considerations around privacy and ethical implications in public settings.

One common way to address this issue is using post-processing techniques that mask out the sensitive content after collection Morales et al. (2019a). However, this approach is risky because raw data remains exposed to theft or misuse before processing.

1.2 Objectives and Contributions

To address this issue, we propose an "End-to-End Human-Head Localisation via Body-Only Detection" model. With this model, the robot can infer the location of the heads while only body parts are visible in the image, providing crucial information that assists the robot system to protect the individual's privacy, as shown in Figure 2.

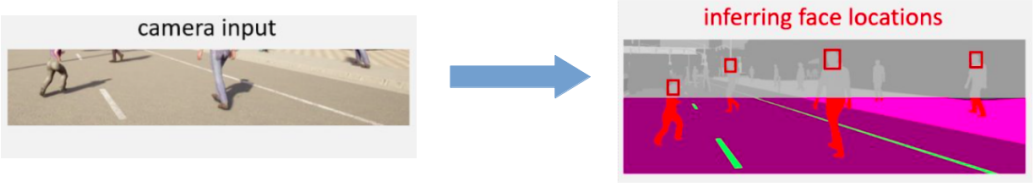


Figure 2: The left image provides an example of raw input for the robot's perception system, showing only portions of the human body are visible at the moment. The right image illustrates the expected output of our model, where the head location is inferred from the raw input. This prediction can potentially be supplied to the robot camera control system, helping to protect the privacy-sensitive information during navigation.

By focusing on body detection rather than post-processing on facial information after collection, we provide a privacy-aware solution that minimises personal data exposure in the first place.

This paper presents the following contributions:

- A novel body-based detection model for head localisation that preserves privacy.
- Implementation of two architectures: a YOLO11-Pose-based model and a ResNet-based model to compare performance.
- Evaluation on the CrowdHuman dataset to demonstrate the effectiveness of our approach in diverse and crowded scenarios.

Our work emphasises the importance of privacy-aware robotics and aims to set a foundation for future research in developing ethical robot perception systems.

2 Related Work

2.1 Privacy-Aware Perception Systems

2.1.1 Privacy-Aware Perception Systems in Robotics

As robots become increasingly present in human environments, they must address the privacy implications of collecting and processing sensitive data, such as facial information, to gain user trust and societal acceptance. Robots need to adhere to legal, ethical, and contextual norms when handling personal data, especially in scenarios involving multiple users or bystanders (Levinson et al., 2024). This necessitates the development of privacy-aware perception systems that ensure responsible data management while maintaining effective robot operation.

2.1.2 Head Inpainting for Privacy-Preserving Visual Perception

This approach uses head inpainting to replace visible heads with anonymized, realistic representations Sun et al. (2018) . It involves generating facial landmarks from body pose and then using these landmarks for inpainting. This technique preserves privacy by obfuscating identifiable features while maintaining scene coherence, allowing robots to adhere to privacy norms effectively.

2.1.3 Detecting and Blurring Potentially Sensitive Personal Information Containers in Images

This model uses obfuscation as a method to ensure privacy by masking sensitive information, like faces or license plates, in the visual input collected by a robot before processing. The obfuscation is achieved through various image-blurring techniques, which minimise the risk of exposing private data while still enabling the robot to use the images for navigation Morales et al. (2019b).

2.2 Head and Body Detection Models

2.2.1 OpenPose for Head and Body Detection

OpenPose (Cao et al., 2019) is a real-time multi-person pose estimation framework that detects 2D keypoints for body, hand, foot, and face. It uses Part Affinity Fields (PAFs) to associate detected keypoints across body parts, enabling accurate head and body localisation. As a bottom-up approach, OpenPose avoids early detection failures typical of top-down methods, ensuring robust pose estimation even in crowded scenarios. OpenPose's ability to detect body parts without relying heavily on facial keypoints aligns with the privacy goals of this report, facilitating body-only head localisation to protect personal data.

2.2.2 YOLO11 for Head and Body Detection Models

YOLO11 is the latest iteration in the YOLO (You Only Look Once) series, known for its real-time object detection capabilities (Jocher and Qiu, 2024). It offers significant advancements in feature extraction and processing speed, enabling efficient human pose estimation through its YOLO11-pose variant. With support for detecting key body parts, such as heads and limbs, YOLO11 is well-suited for applications requiring precise head and body localisation. Its optimised architecture ensures higher accuracy with fewer parameters, making it effective for both cloud-based and edge device deployments. YOLO11's adaptability across environments makes it a robust choice for privacy-aware perception tasks.

Despite existing advanced techniques, there is no ideal method currently allows a robot to have a content-aware perception system that helps robot to proactively protect privacy. Motivated by this, we developed an end-to-end model that indirectly identifies privacy-sensitive content in perception, allowing the robot to avoid collecting such data in the first place.

3 Methodology

3.1 Dataset and Pre-processing

3.1.1 Dataset

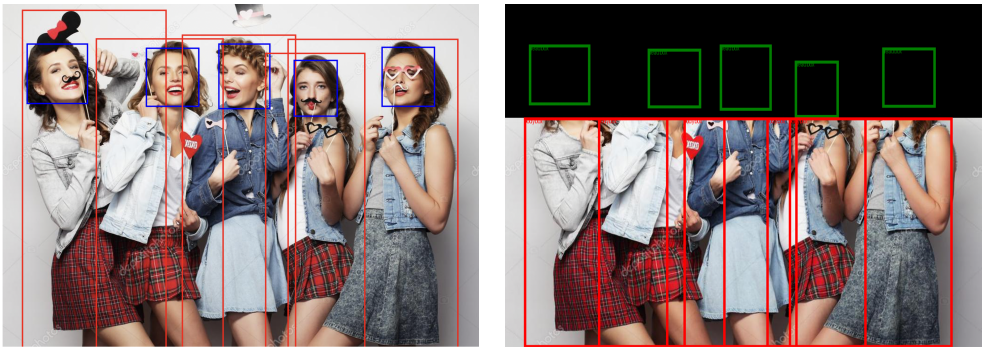
For this project, we use the CrowdHuman dataset, which is a benchmark dataset designed to evaluate detection models in crowded scenarios.Shao et al. (2018) CrowdHuman contains:

- 15,000 images for training,
- 4,370 imagefor validation, and
- 5,000 images for testing.

Each human instance within the dataset is annotated with 3 bounding boxes (Bbox):

- A head Bbox,
- A visible-region Bbox, and
- A full-body Bbox.

These annotations make the dataset particularly suitable for developing and evaluating models focused on human detection tasks, including head localisation from body cues. An example is shown in Figure 3, we only use the body and head Bbox in our project.



(a) (b)

Figure 3: Example image before and after pre-processing. (a) is the original image before pre-processing, and (b) is the image after pre-processing.

3.1.2 Pre-processing Steps

To ensure the dataset aligns with the requirements of our body-only head localisation model, the following pre-processing steps are applied:

1. Filtering Images:
   - We filter out images that contain 7 or fewer persons for simplicity of our tasks.
2. Masking Regions:
   - All areas above the bottom edge of the lowest head Bbox in each image are masked in black, ensuring that only body parts are visible. This step helps maintain a privacy-aware dataset by excluding facial regions.
3. Resizing and Boundary Checking:
   - Each image is resized to 960x576, and the Bbox labels are scaled accordingly. We also perform boundary checks to ensure that the ground truth coordinates remain valid throughout the process.

This pre-processing pipeline ensures the input data is optimised for training the head localisation model, focusing on body detection while minimising facial exposure to align with the privacy-preserving goals of our system. Right image in Figure 3 is an example image after the data pre-processing.

3.2 Model Architectures

Our model architecture is a deep learning-based system that integrates two pre-trained models for feature extraction from images, as illustrated in Figure 4 below. The input image is initially processed by a pre-trained YOLO 11 model Khanam and Hussain (2024), which generates predicted body bounding boxes (Bbox) for human bodies within the image. Only predicted Bboxes with an Intersection over Union (IoU) greater than 0.5 compared to the ground truth Bbox are retained for further processing. The individual body images i_b are then obtained by cropping the original image according to the retained Bbox. Subsequently, we use a pre-trained YOLO 11-Pose model Khanam and Hussain (2024) to capture the pose keypoints from each i_b. These keypoints are then concatenated with the corresponding body Bbox for each detected individual, forming a combined feature set. This combined feature is then fed into a head localisation multi-layer perceptron, which predicts the head Bbox for each detected body.

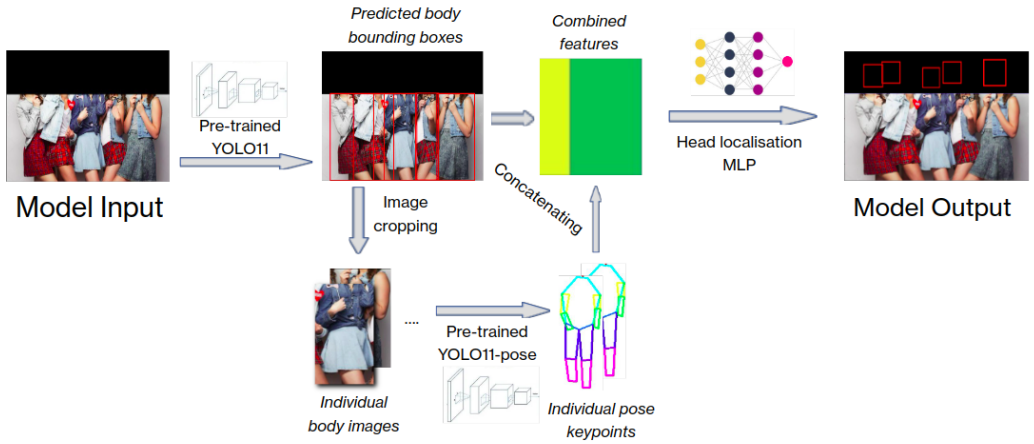
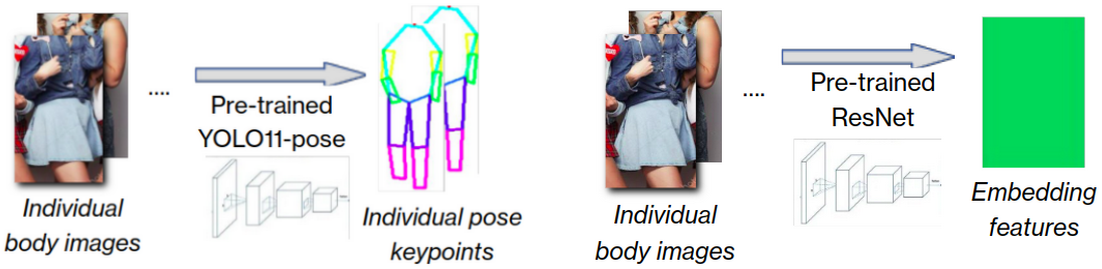


Figure 4: An overview of the model, it takes image as input where only human bodies are visible, and predicts the location of the head using combined individual pose keypoints and body Bbox features.

This model is what we called "YOLO 11-Pose-based model", we also developed another model called "ResNet-based model" for result comparison. The only difference is the YOLO 11-Pose-based model extracts the keypoints from each i_b, where the ResNet-based model uses ResNet He et al. (2015) to extract the embedding features from each i_b for subsequent tasks. As demonstrated in Figure 5 below.



(a) (b)

Figure 5: Image (a) - Structure adopted by YOLO 11-Pose-based model. Image (b) - Structure adopted by ResNet-based model.

3.3 Training Procedure

We train both YOLO 11-Pose-based and ResNet-based models on one GeForce RTX 3090 GPU. We use the Adam optimizer Kingma and Ba (2017) with a learning rate of 1×10^{-3} for all experiments. We train all models with batch size 32 for 15 epochs.

For head Bbox regression, we implement the Smooth L1 Loss (also known as the Huber Loss) for our model. Let $\mathbf{y} = (x_1, y_1, x_2, y_2)$ represent the ground-truth Bbox coordinates (top-left (x_1, y_1) and bottom-right (x_2, y_2) points), and $\hat{\mathbf{y}} = (\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2)$ represent the predicted Bbox coordinates.

The Smooth L1 Loss for each Bbox component is calculated as:

$$\text{SmoothL1}(d) = \begin{cases} 0.5\, d^2 & \text{if } |d| < 1 \\ |d| - 0.5 & \text{otherwise} \end{cases}$$

where d is the difference of one component between predicted and ground-truth Bbox, i.e., $d = \hat{y}_i - y_i$.

The total Smooth L1 Loss for each head Bbox regression is:

$$\mathcal{L}_{\text{h-bbox}} = \sum_{d_i \in \{x_1 - \hat{x}_1, y_1 - \hat{y}_1, x_2 - \hat{x}_2, y_2 - \hat{y}_2\}} \text{SmoothL1}(d_i)$$

4 Experiments and Results

4.1 Experimental Setup

In this section, we describe the experimental setup used to evaluate the performance of the YOLO11-pose-based and ResNet-based models for head localisation. Both models were trained on the pre-processed CrowdHuman dataset. The IoU threshold was set to 0.20, where a predicted Bbox is considered correct if its IoU with the ground-truth Bbox is greater than or equal to 0.20. Figure 6 presents an example comparing predicted Bboxes with ground truth Bboxes using IoU values.

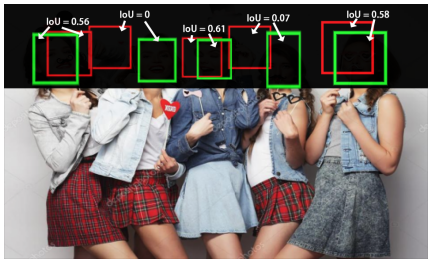


Figure 6: Example output of IoU values between predicted (red) and ground-truth (green) boxes, illustrating varying overlap: well-aligned (e.g., IoU = 0.61) and minimal or none (e.g., IoU = 0.07, IoU = 0).

4.2 Evaluation Metrics

We use the following metrics to evaluate model performance:

- **Precision**: Measures the proportion of correctly predicted heads among all detected heads.

$$\text{Precision} = \frac{TP}{TP + FP}$$

  A higher precision indicates fewer false positives (FP), which is crucial in privacy-sensitive scenarios.

- **Recall**: Measures the proportion of true heads that the model successfully detected.

$$\text{Recall} = \frac{TP}{TP + FN}$$

  A higher recall ensures fewer missed heads, which is important in crowded environments.

- **AP (Average Precision)**: Summarises the precision-recall trade-off across different IoU thresholds. AP reflects overall detection accuracy.
- **F1 Score**: The harmonic mean of precision and recall, balancing both metrics.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.3 Quantitative Results

Model	Precision	Recall	AP	F1 Score
ResNet-based	0.38	0.60	0.23	0.46
YOLO11-Pose-based	0.63	0.71	0.45	0.67

Table 1: Quantitative results of the ResNet-based and YOLO11-pose-based models.

Table 1 summarises the performance comparison between the YOLO11-pose-based and ResNet-based models. The YOLO11-pose-based model achieves higher scores across all metrics, demonstrating better overall accuracy, with fewer false positives and missed detections than the ResNet-based model. With a precision of 0.63 and a recall of 0.71, the YOLO11-pose-based model balances accurate predictions and low false positives. Its higher F1 score (0.67) reflects better performance in both precision and recall compared to the ResNet model's F1 score of 0.46.

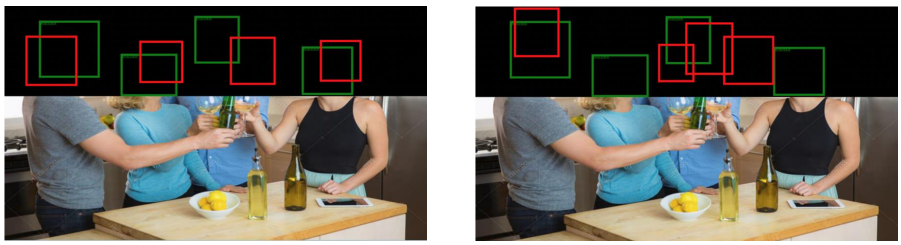
IoU Threshold	Precision	Recall	AP
0.10	0.68	0.75	0.51
0.15	0.64	0.72	0.48
0.20	0.63	0.71	0.45

Table 2: Performance of the YOLO11-pose-based model at different IoU thresholds.

As shown in Table 2 above, which presents the YOLO11-pose-based model's performance evaluated at different IoU thresholds (0.10, 0.15, and 0.20). As the IoU threshold decreases, AP, precision, and recall scores improve, showing how performance varies with different localisation requirements. This trend highlights how relaxing localisation requirements can improve precision, recall, and AP scores. However, stricter thresholds result in slightly lower performance, reflecting the trade-off between overlap criteria and detection accuracy.

4.4 Qualitative Results

Figure 7 presents examples of model outputs for both the YOLO11-pose-based and ResNet-based models. Green boxes represent the ground-truth labels, and red boxes indicate the model's predictions. In these examples, the YOLO11-pose-based model provides more accurate predictions, with Bboxes closely aligned with the ground-truth labels.



(a) (b)

Figure 7: Qualitative results comparing the YOLO11-pose-based and ResNet-based models. (a) shows the output of the YOLO11-pose-based model, with red bounding boxes closely aligning with the green ground-truth labels. (b) shows the predictions of the ResNet-based model, which struggles in complex scenarios, resulting in more deviations between predictions and ground-truth labels.

5 Discussion

5.1 Summary of Findings

Our project presents a privacy-aware perception system that effectively localises human heads using body-only detection, which can potentially help prevent the collection of personal facial data by robot perception system.

Our experiments indicate that the YOLO11-pose-based model has better performance compare to the ResNet-based model across key metrics, such as Precision, Recall, AP, and F1 Score. Where the YOLO11-pose-based model achieved a 63% Precision and a 71% Recall, demonstrating its suitability for scenarios requiring high sensitivity to head localisation without facial recognition. Additionally, varying the IoU threshold illustrated a performance trade-off, where lower thresholds increased recall and AP, allowing flexibility for different application requirements. Qualitative results further underscored the YOLO11-pose model's robustness, as it maintained accuracy in crowded and complex scenes, essential for public and privacy-sensitive environments.

5.2 Limitations of the Study

Although our proposed model shows promising results, there are several limitations to address. One major limitation is the model's dependence on the quality of the prediceted body Bboxes. If the detected body regions are inaccurate or incomplete, the head localisation performance deteriorates significantly.

Additionally, the dataset filtering process limits the model's exposure to complex and crowded scenarios. Currently, images with fewer than a specific number of people are excluded, restricting the model's ability to handle densely populated environments with overlapping bodies. This limitation reduces the model's generalisability to real-world applications, such as crowded public spaces.

The two pre-trained models (YOLO11/YOLO11-pose and ResNet) were set in eval model during training, which means the models didn't fine-tune for this specific head inference task, which could potentially be a reason that decreased the model's performance.

5.3 Future Work

5.3.1 Data Collection

One key improvement is to increase the maximum number of people in the images in the dataset filtering process. By increasing it, the model will be exposed to more complex and realistic situations, such as densely populated scenes. This adjustment will help the model learn to perform better in scenarios with significant occlusion and overlapping bodies, ultimately improving its robustness. With a broader variety of training instances, the system could potentially generalise to real-world applications, including crowded public spaces like train stations, events, and markets.

5.3.2 Model Improvements

In addition to enhancing the dataset, further improvements can be made to the model architecture. One approach is to explore alternative pose estimation models that may offer more precise and discriminative keypoints feature. Using advanced pose estimators could improve the quality of body information fed into the head localisation model, leading to better predictions.

Another promising direction is to fine-tune the YOLO11 model to focus on detecting body part without including head information. This improvement would significantly reduce the model's reliance on facial features, enhancing its performance on body detection for this specific task.

This direction makes the model not only suitable for robot privacy-aware perception system, but also for other privacy-sensitive applications, such as home assistant or healthcare settings, while maintaining robust detection capabilities even in complex environments. These adjustments will ensure the model achieves higher reliability and broader usability without compromising on its privacy-preserving objectives.

References

CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; AND SHEIKH, Y., 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. https://arxiv.org/abs/1812.08008.

GIRASA, R., 2020. *Ethics and Privacy I: Facial Recognition and Robotics*, 105–146. Springer International Publishing, Cham. ISBN 978-3-030-35975-1. doi:10.1007/978-3-030-35975-1_4. https://doi.org/10.1007/978-3-030-35975-1_4.

HE, K.; ZHANG, X.; REN, S.; AND SUN, J., 2015. Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385.

JOCHER, G. AND QIU, J., 2024. Ultralytics yolo11. https://github.com/ultralytics/ultralytics.

KHANAM, R. AND HUSSAIN, M., 2024. Yolov11: An overview of the key architectural enhancements. https://arxiv.org/abs/2410.17725.

KINGMA, D. P. AND BA, J., 2017. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980.

KIRSCHGENS, L. A.; UGARTE, I. Z.; URIARTE, E. G.; ROSAS, A. M.; AND VILCHES, V. M., 2021. Robot hazards: from safety to security. https://arxiv.org/abs/1806.06681.

LEVINSON, L.; DIETRICH, M.; SARKISIAN, A.; SABANOVIC, S.; AND SMART, W. D., 2024. Privacy aware robotics. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24 (Boulder, CO, USA, 2024), 1335–1337. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3610978.3638161. https://doi.org/10.1145/3610978.3638161.

MORALES, J.; HULIGANGA, V.; PASAOA, J.; AND MELAD, N., 2019a. *Detecting and Blurring Potentially Sensitive Personal Information Containers in Images Using Faster R-CNN Object Detection Model with TensorFlow and OpenCV*. Ph.D. thesis.

MORALES, J.; HULIGANGA, V.; PASAOA, J.; AND MELAD, N., 2019b. *Detecting and Blurring Potentially Sensitive Personal Information Containers in Images Using Faster R-CNN Object Detection Model with TensorFlow and OpenCV*. Ph.D. thesis.

SHAO, S.; ZHAO, Z.; LI, B.; XIAO, T.; YU, G.; ZHANG, X.; AND SUN, J., 2018. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, (2018).

SUN, Q.; MA, L.; OH, S. J.; GOOL, L. V.; SCHIELE, B.; AND FRITZ, M., 2018. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5050–5057.

6 Declarations

Chenyang Li, u7631563. I contribute to the project by developing the idea, researching the dataset, writing model, training, and testing code, running experiments, preparing the video, writing and revising the report.

Yao Chen, u7722352. I contribute to the project by collecting the dataset, doing pre-processing to the original dataset, training the model, preparing the video, writing the report.

Qiance Zhou, u7520051. I contribute to the project by finding and testing the pre-trained models, organising the experiment results, preparing the video, and writing the report draft.

Baek Cheol Kim, u7617018. I contribute to the project by pre-processing dataset, running and training the model with GPU cluster server, preparing the video, writing the report.