AnonymEyes: Lightweight and Affordable Privacy-Enhanced Human Detection via Depth Sensing on Edge Devices

Sercan Yeşilköy[1], Mohsen Ali Alawami[2], Yoon-Ho Choi[1]
[1]*School of Computer Science and Engineering*, *Pusan National University*, Busan, Republic of Korea
[2]*Division of Computer Engineering*, *Hankuk University of Foreign Studies*, Yongin-si, Republic of Korea
sercanyesilkoy@pusan.ac.kr, mohsencomm@hufs.ac.kr, yhchoi@pusan.ac.kr

***Abstract*—Privacy concerns in human monitoring systems have become increasingly critical as surveillance technologies proliferate in public and private spaces. Conventional detection solutions rely on RGB-based systems, raising serious privacy concerns due to potential identity disclosure. This paper introduces AnonymEyes, a lightweight, affordable, and privacy-enhanced human detection system utilizing Time-of-Flight (ToF) depth cameras with edge computing devices such as the Raspberry Pi. By exclusively leveraging depth data while discarding RGB-based data, AnonymEyes enhances individual anonymity while maintaining high detection performance. The system employs YOLOv8 Nano architecture based framework to process ToF depth frames, achieving real-time detection on resource-constrained environments. The system was trained and validated on over 1,500 annotated ToF depth frames encompassing diverse poses, clothing, and appearances. The system demonstrates exceptional metrics: 99.90% precision, 100.00% recall, 99.95% F1-score, 99.50% mAP@0.5, and 92.50% mAP@0.5:0.95 across stricter IoU thresholds. AnonymEyes represents a robust and scalable solution for applications requiring accurate, real-time human presence monitoring in privacy-sensitive, resource-constrained environments.**

***Index Terms*—Privacy-enhanced, Human Detection, Time-of-Flight (ToF) camera, Depth Sensing, Edge Computing, Deep-Learning, YOLOv8, Resource-constrained devices**

I. Introduction

Human detection systems have become increasingly prevalent across diverse applications, from smart building automation and security monitoring to occupancy sensing and space management. However, conventional solutions predominantly rely on RGB-based systems, raising significant privacy concerns due to identity disclosure. Traditional RGB-based detection systems, while offering high accuracy, inherently compromise individual privacy by capturing detailed visual data that can reveal identities. Additionally, RGB processing requires substantial computational resources and higher costs, reducing scalability for large-scale deployments. These limitations significantly restrict the deployment of monitoring technologies, particularly in sensitive environments such as healthcare facilities, private offices, and residential spaces where privacy expectations are paramount.

In response to these privacy challenges, this study introduces AnonymEyes, a privacy-enhanced human detection system leveraging Time-of-Flight (ToF) depth cameras (shown in Fig. 1) with affordable edge computing devices such as the Raspberry Pi. Through exclusive reliance on depth data and complete omission of RGB inputs, the system enhances individual anonymity while maintaining robust detection performance, effectively mitigating key privacy concerns associated with facial recognition and visual identification. The system

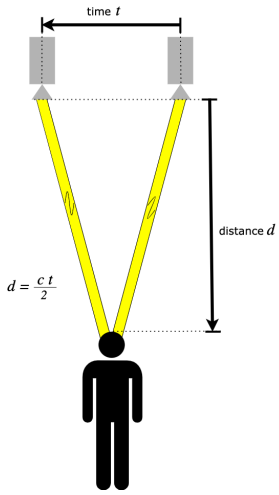

Fig. 1. Demonstration of the Time-of-Flight (ToF) camera.The camera emits infrared light toward the scene and measures the round-trip travel time of the reflected signal. The depth d is computed from the measured delay t as $d = \frac{ct}{2}$, where c denotes the speed of light.

employs a lightweight YOLOv8 Nano architecture trained on ToF depth frames, enabling real-time detection on resource-constrained edge devices. All detection operations occur locally, eliminating cloud-based computation or external data transmission beyond initial model training and further minimizing privacy concerns. To ensure robust performance across diverse real-world scenarios, we developed a comprehensive dataset of over 1,500 annotated ToF depth frames captured from individuals with varying poses, clothing, heights, and orientations. This dataset provides essential variations that improve the system's generalization and reliability under practical deployment conditions. This paper presents the complete system pipeline from ToF camera calibration through depth data acquisition, spatial transformation, and YOLO-based human detection, including dataset development, edge implementation on resource-constrained devices, and comprehensive

validation. Our results demonstrate that human detection using ToF depth data can achieves exceptional accuracy metrics while eliminating facial recognition capabilities and reducing privacy invasion compared to RGB-based approaches. The system provides a scalable, cost-effective solution (under $150) for applications requiring reliable human presence monitoring, with fully local processing on edge devices eliminating cloud dependency and associated privacy risks. The key contribution of our system can be summarized as below:

- **Mitigating Privacy Risk by-Design:** An end-to-end ToF depth-based human detection framework that mitigates visual privacy risks by eliminating RGB data capture while achieving exceptional performance (99.90% precision, 100.00% recall, 99.95% F1-score, 99.50% mAP@0.5).
- **Edge-Optimized Framework:** Tailored YOLOv8 Nano based framework for real-time human detection on resource-constrained devices with fully on-device processing, enhancing the privacy mitigating deployment without cloud infrastructure.
- **Comprehensive ToF Depth Dataset:** Development and annotation of over 1,500 ToF depth frames encompassing diverse poses, clothing, heights, and orientations, providing a valuable resource for privacy-enhanced human detection research.
- **Cost-Effective Deployment Solution:** Demonstration of a practical, scalable system costing under $150 that achieves exceptional metrics suitable for privacy-sensitive applications in healthcare, smart homes, and office environments.

II. Related Works

Human detection research spans smart building automation, occupancy monitoring, security systems, and human-computer interaction. Early approaches used classical computer vision [1] and handcrafted features like HOG with SVMs [2], facing limitations under occlusion and challenging lighting. Deep learning approaches with CNNs [3], [4] and real-time frameworks like YOLO and SSD [5]–[7] improved accuracy but raised significant privacy concerns due to RGB imagery. Alternative modalities include thermal imaging [8], LiDAR with 3D CNNs or SVMs [9]–[11], and ToF depth sensors [12], [13]. Luna et al. [14] achieved ∼150 fps using overhead ToF RGB-D cameras, while Wang et al. [15] achieved 97.73% accuracy with morphological classification. However, RGB-D cameras compromise privacy through RGB capture, and ToF systems face noise from reflective surfaces and occlusions. While [16] achieved good segmentation performance with hemisphere LiDAR, the high cost of such systems raises scalability concerns for widespread deployment. Edge implementations show promise [17], but RGB-based [18] and RGB-D approaches [19] lack privacy preservation, while lightweight CNNs [20] may have insufficient frame rates. These limitations motivate our depth-only ToF approach on low-cost edge devices, eliminating RGB data entirely while achieving real-time performance

III. Proposed System

This section describes the design and implementation of the proposed **AnonymEyes** system a privacy-enhanced, affordable, and lightweight framework for real-time human detection. The system leverages a Time-of-Flight (ToF) depth camera in combination with an edge computing device, the Raspberry Pi 5, to detect human presence using only depth information. By operating entirely on depth data, stored in NumPy array format, the system inherently prevents the capture of personally identifiable visual features, enhancing privacy.

Fig. 2. The Arducam ToF camera connects via CSI to the Raspberry Pi Model 5, serving as the central computing resource for our Edge-IoT setup.

A. System Design

The hardware configuration (see Fig. 2) employs an Arducam ToF camera mounted in an overhead position to capture depth arrays covering the monitored area (see Fig. 4). ToF sensing operates by emitting modulated infrared light pulses and measuring the phase shift or return time from scene surfaces as reviewed on. [21]. This produces dense depth maps, unaffected by ambient lighting or surface color variations. The system architecture comprises four primary functional components:

1) **Field-of-View calibration, Focal Length Computation and Depth data acquisition**: Continuous acquisition of depth arrays from the ToF camera with concurrent field-of-view calibration and focal length computation to ensure accurate spatial measurements.
2) **Adaptive Kalman-filter-based depth processing**: Implementation of advanced state-space filtering techniques to reduce temporal noise and enhance depth data quality through predictive smoothing algorithms.
3) **Depth-to-image conversion**: Transformation of filtered depth maps from NumPy arrays into single-channel grayscale images with normalized intensity values, ensuring compatibility with the YOLOv8 Nano architecture while preserving spatial geometry.
4) **Human detection**: YOLOv8 Nano-based object detection pipeline specifically trained for human identification in depth imagery, providing real-time inference capabilities.

As depicted in Fig. 3, these components form a sequential processing pipeline from raw depth acquisition through noise reduction, data formatting, and detection stages. This modular design enables independent optimization of each component while maintaining privacy through complete exclusion of RGB data.

All computations, excluding the YOLOv8 Nano training phase, are performed entirely on the edge device, eliminating any reliance on external servers or cloud services. After offline model training and deployment, the system functions autonomously without network connectivity, thereby ensuring that all depth data remains strictly local to the device and never transmitted externally.

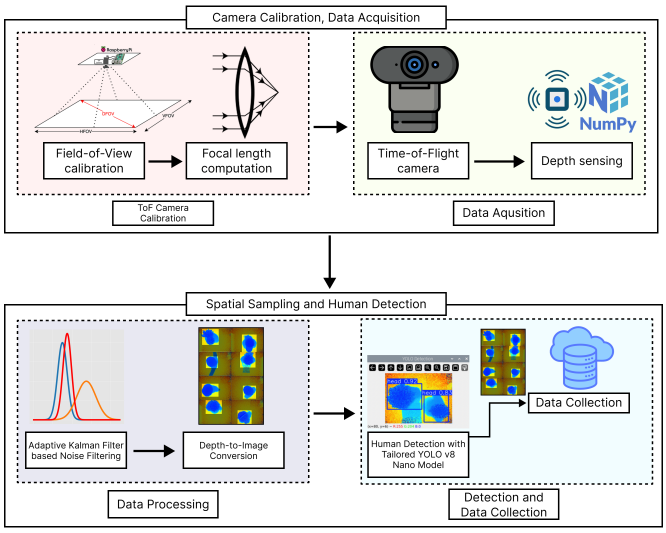

Fig. 3. The Overview of AnonymEyes System.

B. Methodology

1) Field-of-View Calibration and Focal Length Computation: Accurate 3D spatial mapping requires precise camera calibration parameters. The specifications of the Arducam ToF camera provided [22]:

$$\theta_h = 62.8^\circ \quad \text{(horizontal field of view)} \tag{1}$$

$$\theta_v = 37.9^\circ \quad \text{(vertical field of view)} \tag{2}$$

where θ_h denotes the horizontal field-of-view angle in degrees, and θ_v represents the vertical field-of-view angle in degrees.

These field-of-view angles (see Fig. 4) are converted to focal lengths in pixels using the pinhole camera model [23]:

$$f_x = \frac{W/2}{\tan(\theta_h/2)} \tag{3}$$

$$f_y = \frac{H/2}{\tan(\theta_v/2)} \tag{4}$$

where f_x and f_y are the focal lengths in pixels along the horizontal and vertical axes, respectively, W represents the image width in pixels, and H represents the image height in pixels.

The principal point is assumed to be at the image center:

$$(c_x, c_y) = \left(\frac{W}{2}, \frac{H}{2}\right) \tag{5}$$

where c_x and c_y denote the x and y coordinates of the principal point (optical center) in the image coordinate system, measured in pixels.

2) Adaptive Kalman-Filtered Depth Frame Processing: Time-of-Flight (ToF) cameras inherently produce noisy depth measurements due to sensor limitations and environmental factors. To obtain stable and reliable depth readings while maintaining computational efficiency, the system employs an adaptive Kalman filtering approach that estimates noise characteristics from strategic keypoints and applies uniform temporal filtering to the entire depth frame.

Three strategic keypoints along the horizontal centerline provide real-time noise characterization:

$$\text{Left: } (u_L, v_c) = (0, H/2) \tag{6}$$

$$\text{Center: } (u_C, v_c) = (W/2, H/2) \tag{7}$$

$$\text{Right: } (u_R, v_c) = (W-1, H/2) \tag{8}$$

where u_L, u_C, and u_R denote the horizontal pixel coordinates of the left, center, and right keypoints respectively, and v_c represents the vertical centerline coordinate.

Each keypoint maintains an independent Kalman filter tracking depth measurements over time. The measurement noise variance is estimated as the average across the three keypoints:

$$\sigma^2_{\text{meas}} = \frac{1}{3} \sum_{i \in \{L,C,R\}} \sigma_i^2(t) \tag{9}$$

where $\sigma_i^2(t)$ represents the estimated measurement noise variance at keypoint i at time t.

Using these adaptively estimated noise parameters, a frame-level Kalman filter is applied uniformly across all pixels. For each pixel (u, v), the filter maintains two states:

- **Depth estimate**: $d_t^{(u,v)}$ — the filtered depth value at pixel (u, v) at time t, measured in millimeters
- **Depth velocity**: $v_t^{(u,v)}$ — the temporal rate of change of depth at pixel (u, v), measured in millimeters per frame

At each time step, the filter performs two operations for every pixel:

(i) Prediction Step: The predicted depth $\hat{d}_t^{(u,v)}$ is estimated from the previous depth and velocity:

$$\hat{d}_t^{(u,v)} = d_{t-1}^{(u,v)} + v_{t-1}^{(u,v)} \cdot \Delta t \tag{10}$$

where Δt denotes the time interval between consecutive depth frame acquisitions, determined by the ToF camera's frame rate.

(ii) Correction Step: The filtered depth $d_t^{(u,v)}$ is obtained by blending the prediction with the new measurement $z_t^{(u,v)}$:

$$d_t^{(u,v)} = \hat{d}_t^{(u,v)} + K_t \left(z_t^{(u,v)} - \hat{d}_t^{(u,v)} \right) \tag{11}$$

where $z_t^{(u,v)}$ is the raw depth measurement from the ToF sensor at pixel (u, v) at time t, and K_t is the Kalman gain computed using the shared measurement noise variance σ^2_{meas} across all pixels. The Kalman gain automatically adapts based on the estimated measurement noise and configured process noise, ensuring responsiveness to actual depth changes while suppressing temporal fluctuations.

This adaptive approach provides several advantages: (1) computational efficiency by estimating noise from only three keypoints rather than analyzing all pixels independently, (2) spatial awareness by sampling diverse sensor regions across the field of view, and (3) automatic adaptation to changing environmental conditions affecting sensor noise characteristics. The filtered depth frame $D_t = \{d_t^{(u,v)} | \forall (u, v)\}$ provides temporally stable measurements that preserve spatial geometry while suppressing sensor noise, creating a robust foundation for the subsequent depth-to-image conversion and human detection stages.

3) Depth-to-Image Conversion: To prepare depth data for the YOLOv8 Nano detection network, filtered depth maps from the Adaptive Kalman-Filter processing stage are converted into single-channel grayscale images. The conversion leverages OpenCV (Open Source Computer Vision Library) to transform NumPy array-based depth representations into image format, with depth values normalized to the 0-255 intensity range. This standardized representation preserves spatial geometry and depth discontinuities while ensuring compatibility with the YOLO architecture and enabling efficient real-time inference.

4) Human Detection: The detection module employs the YOLOv8 Nano architecture, specifically trained for processing single-channel depth imagery. The model takes preprocessed depth frames as input and outputs bounding box predictions with associated confidence scores.

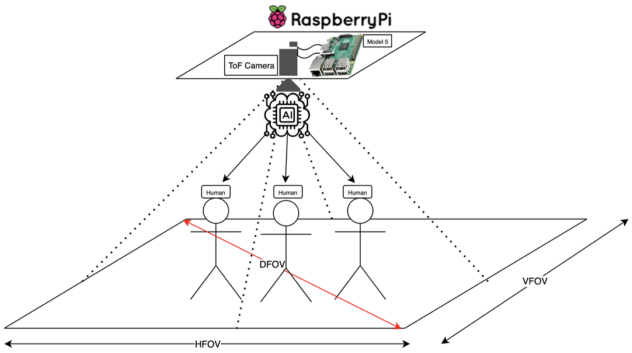

Fig. 4. Human detection process using the overhead ToF camera system. The AI module processes depth data to detect multiple human subjects within the field of view, demonstrating the system's capability for real-time human presence detection while maintaining privacy through depth-only sensing. HFOV, VFOV, and DFOV represent horizontal, vertical, and diagonal field-of-view angles, respectively, defining the ToF camera's coverage area.

The training dataset comprises over 1,500 annotated ToF depth frames, encompassing diverse poses, clothing types, and appearances to ensure robust performance under real-world variability. The model learns to recognize human silhouettes and depth discontinuities rather than texture or color cues, enhancing privacy while retaining robustness.

C. Dataset and Model Training

The dataset captures realistic variations in appearance and positioning, enabling generalization across typical human

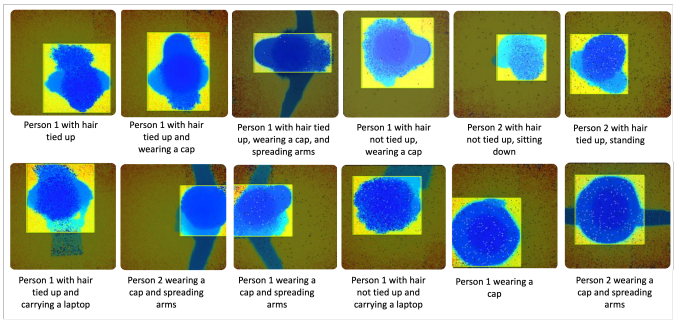

Fig. 5. An example frame from our dataset, which contains 1501 frames with different scenarios such as different outfits, positions, and accessories like caps, headset, backpacks, etc.

monitoring scenarios. Fig. 5 shows example frames with different positions, outfits, postures, and accessories. The dataset comprises 1,501 frames from 2 unique subjects, split into training (70%, 1,051 frames), validation (15%, 225 frames), and test (15%, 225 frames) sets with subject-independent partitioning to ensure unbiased evaluation. Despite the limited number of subjects, the diversity in appearance and pose supports effective training for depth-based detection.

The model employs the YOLOv8 Nano architecture [24], a lightweight YOLO variant optimized for resource-constrained devices, trained on the ToF depth frames with diverse appearances and poses. Training was conducted on NVIDIA Tesla T4 GPU workstation using data augmentation (horizontal flipping $p = 0.5$, mosaic augmentation, and `albumentations` including `Blur`, `MedianBlur`, `ToGray`, and `CLAHE`), the AdamW optimizer [25] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay=0.0005) with a base learning rate of 0.002 and a cosine annealing scheduler, batch size of 16, input resolution of 640×640 pixels, and early stopping (patience=25 epochs, monitoring validation mAP@0.5) to prevent overfitting, converging at epoch 98.

IV. Experimental Results and Analysis

This section evaluates the AnonymEyes system across three dimensions: real-time system performance on edge devices, YOLOv8 Nano model accuracy using precision, recall, F1-score, and mAP metrics, and system scalability. The results demonstrate that AnonymEyes achieves superior performance while mitigating privacy risks, positioning it as a viable alternative for privacy-sensitive applications.

A. System Performance

The AnonymEyes system demonstrates robust real-time performance on the Raspberry Pi 5 edge computing platform. The complete processing pipeline achieves total latency:

$$t_{\text{total}} = t_{\text{acquisition}} + t_{\text{kalman}} + t_{\text{conversion}} + t_{\text{detection}} \quad (12)$$

where t_{total} represents total frame processing time, with each component t denoting processing time for depth data acquisition, Kalman filtering, depth-to-image conversion, and YOLOv8 detection. This achieves approximately 15 frames

TABLE I
REPORTED PERFORMANCE COMPARISON WITH EXISTING WORKS

Study	Metrics	System Cost	Model and Method
Luna et al. 2016 [14]	F1 Score: 99.57% Recall: 99.57% Precision: –.–% FPS: $\sim$ 150 FPS	At least over 1000 USD	Local Maxima Detection, ROI Estimation, PCA-Based Classifier
Wang et al. 2019 [15]	F1 Score: 95.04% Recall: 100.00% Precision: 90.55% FPS: $\sim$ 40 FPS	At least over 1500 USD	Local Pooling Maxima Search Shallow CNN Classification
Seliunina et al. 2025 [16]	F1 Score: 97.00% Recall: –.–% Precision: 98.00% FPS: $\sim$ – FPS	At least over 16000 USD	Multi-Channel LiDAR Processing, Positional Encoding, MaskDINO-Based Human Detection
AnonymEyes	F1 Score: **99.95%** Recall: **100.%** Precision: **99.90%** mAP@0.5: **99.50%** mAP@0.5:0.95: **92.50%** FPS: 15 FPS	**148 USD**	Adaptive Kalman-filtered based depth processing, YOLOv8 Nano, Edge Computing

per second (FPS) with $t_{\text{total}} < 100$ms per frame, sufficient for real-time human presence monitoring applications.

Memory utilization remains below 2GB during operation, with the YOLOv8 Nano model requiring approximately 6MB of storage. The system maintains stable performance across extended periods without memory leaks or degradation, suitable for continuous deployment.

B. Model Performance

The YOLOv8 Nano model demonstrates exceptional detection performance on the custom ToF depth dataset. Training convergence analysis (Fig. 6) shows steady improvement across all metrics, with the model achieving optimal performance after approximately 100 epochs. The best model checkpoint, selected based on highest validation mAP@0.5, was evaluated on the held-out test set.

Quantitative evaluation on the independent test dataset yields the following comprehensive metrics:

- **Precision**: 99.90% - indicating minimal false positive detections
- **Recall**: 100.00% - demonstrating excellent detection sensitivity
- **F1-Score**: 99.95% - confirming balanced precision-recall performance
- **Mean Average Precision (mAP@0.5)**: 99.50% - validating accurate localization at IoU threshold 0.5
- **Mean Average Precision (mAP@0.5:0.95)**: 92.50% - maintaining high accuracy across stricter IoU thresholds

The training curves indicate robust convergence with total loss decreasing from initial high values to near-zero levels. Box regression loss and classification loss components both demonstrate consistent improvement, reflecting the model's ability to accurately predict bounding boxes and classify human subjects in depth imagery. Evaluation across different poses, clothing variations, and spatial positions confirms the model's generalization capability.

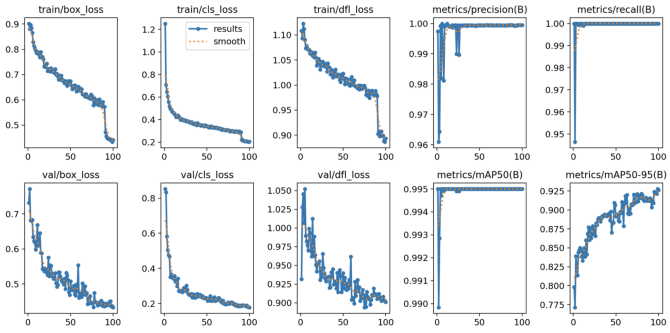

Fig. 6. Training and validation performance curves of the YOLOv8 Nano model, showing loss components and detection metrics (precision, recall, and mAP) over training epochs.

Table I summarizes reported performance of our method against prior works and state-of-the-art methods for contextual comparison. Results demonstrating that AnonymEyes achieves competitive metrics while operating on significantly more cost-effective hardware. The F1-score of 99.95% is higher than the values reported by Luna et al. [14]'s 99.57%, Wang et al. [15]'s 95.04%, and Seliunina et al. [16]'s 97.00%, while a recall of 100% is consistent with the highest recall values reported in prior works [14] [15]. Furthermore, AnonymEyes achieves a precision of 99.90%, which is comparable to the precision reported by Seliunina et al. [16] (98.00%), while maintaining substantially lower system cost (148 USD vs. 16,000 USD) and providing comprehensive mAP metrics across multiple IoU thresholds. Note that, values from Table I for prior works are taken directly from their respective publications; '–' indicates the metric was not reported in the original work.

C. System Scalability

The AnonymEyes system supports scalable deployment from single installations to large-scale distributed networks. At $148 per unit with 8-10 watt power consumption, the low-cost hardware enables cost-effective multi-point scaling. Each unit operates independently with edge-based processing, ensuring linear scaling characteristics without centralized bottlenecks or single points of failure. Moreover, 15 FPS processing adequately covers typical monitoring scenarios [26], with optimization potential through model quantization or hardware acceleration.

V. Limitation and Future Work

Although AnonymEyes achieves strong detection performance on resource-constrained edge devices, several limitations remain. The system currently operates at approximately 15 FPS on a Raspberry Pi 5, which may be insufficient for applications requiring higher temporal resolution; future work will investigate optimization through quantization, and hardware acceleration. Future extensions should explore multi-camera fusion and temporal consistency models to address occlusion, coverage limitations, and detection stability in dynamic scenes.

VI. Conclusion

This paper presents AnonymEyes, a privacy-enhanced human detection system that balances accuracy with privacy in resource-constrained environments. By using ToF depth data instead of RGB imagery, the system provides cost-effective, real-time monitoring with competitive detection performance, while enhancing privacy. With a total cost under $150 and fully edge-based processing, AnonymEyes demonstrates practical scalability for deployment in privacy-sensitive environments such as healthcare facilities, smart homes, and office spaces.

Acknowledgment

This work was supported by the IITP(Institute of Information & Coummunications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2025-RS-2023-00259967) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00217689)

References

[1] J. Zhou and J. Hoang, "Real time robust human detection and tracking system," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*. IEEE, 2005, pp. 149–149.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[3] M. Szarvas, A. Yoshizawa, M. Yamamoto, and J. Ogata, "Pedestrian detection with convolutional neural networks," in *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.* IEEE, 2005, pp. 224–229.

[4] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3258–3265.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] T. Dubail, F. A. Guerrero Peña, H. R. Medeiros, M. Aminbeidokhti, E. Granger, and M. Pedersoli, "Privacy-preserving person detection using low-resolution infrared cameras," in *European Conference on Computer Vision*. Springer, 2022, pp. 689–702.

[9] M. R. Blanch, Z. Li, S. Escalera, and K. Nasrollahi, "Lidar-assisted 3d human detection for video surveillance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 123–131.

[10] A. Günter, S. Böker, M. König, and M. Hoffmann, "Privacy-preserving people detection enabled by solid state lidar," in *2020 16th international conference on intelligent environments (IE)*. IEEE, 2020, pp. 1–4.

[11] L. Wang, *Support vector machines: theory and applications*. Springer Science & Business Media, 2005, vol. 177.

[12] R. Tanner, M. Studer, A. Zanoli, and A. Hartmann, "People detection and tracking with tof sensor," in *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2008, pp. 356–361.

[13] L. Xia, C.-C. Chen, and J. K. Aggarwal, "Human detection using depth information by kinect," in *CVPR 2011 workshops*. IEEE, 2011, pp. 15–22.

[14] C. A. Luna, C. Losada-Gutierrez, D. Fuentes-Jimenez, A. Fernandez-Rincon, M. Mazo, and J. Macias-Guarasa, "Robust people detection using depth information from an overhead time-of-flight camera," *Expert Systems with Applications*, vol. 71, pp. 240–256, 2017.

[15] W. Wang, P. Liu, R. Ying, J. Wang, J. Qian, J. Jia, and J. Gao, "A high-computational efficiency human detection and flow estimation method based on tof measurements," *Sensors*, vol. 19, no. 3, p. 729, 2019.

[16] S. Seliunina, A. Otelepko, R. Memmesheimer, and S. Behnke, "Person segmentation and action classification for multi-channel hemisphere field of view lidar sensors," in *2025 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2025, pp. 817–822.

[17] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Real-time human detection as an edge service enabled by a lightweight cnn," in *2018 IEEE International Conference on Edge Computing (EDGE)*. IEEE, 2018, pp. 125–129.

[18] I. Ahmed, M. Ahmad, J. J. Rodrigues, and G. Jeon, "Edge computing-based person detection system for top view surveillance: Using centernet with transfer learning," *Applied Soft Computing*, vol. 107, p. 107489, 2021.

[19] M. Gochoo, S. A. Rizwan, Y. Y. Ghadi, A. Jalal, and K. Kim, "A systematic deep learning based overhead tracking and counting system using rgb-d remote cameras," *Applied Sciences*, vol. 11, no. 12, p. 5503, 2021.

[20] J. Yrjänäinen, X. Ni, B. Adhikari, and H. Huttunen, "Privacy-aware edge computing system for people tracking," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 2096–2100.

[21] R. Horaud, M. Hansard, G. Evangelidis, and C. Ménier, "An overview of depth cameras and range scanners based on time-of-flight technologies," *Machine vision and applications*, vol. 27, no. 7, pp. 1005–1020, 2016.

[22] Arducam. (2025) Tof camera — specifications. Accessed: 2025-08-12. [Online]. Available: https://docs.arducam.com/Raspberry-Pi-Camera/Tof-camera/TOF-Camera/#specifications

[23] R. Paschotta, "Focal length," *RP Photonics Encyclopedia*, 2007.

[24] M. Yaseen, "What is yolov8: An in-depth exploration of the internal features of the next-generation object detector. arxiv 2024," *arXiv preprint arXiv:2408.15857*, 2024.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[26] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *European conference on computer vision*. Springer, 2020, pp. 474–490.