2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust Person Re-Identification for Service Robots via One-Class Body-Part Transformer and Continual Learning

Enrique Aleman-Gallegos[1] and Sven Wachsmuth[1]

***Abstract*— This work presents a robust person tracking and re-identification system designed for Human-Robot Interaction applications. The approach introduces the One-Class Body-Part (OCBP) Transformer, trained online to model interactions among body-part features and construct a robust target representation. To improve data association and reduce identity swaps during the tracking phase, the SORT tracker is extended with depth information in order to provide correct samples for the Online Continual Learning (OCL) setting. The transformer is further enhanced through the use of pseudo-negative samples, which accelerate convergence during the online learning phase. Ablation studies compare the performance of the memory management system using different sample insertion configurations and highlight the benefit of using pseudo-negative samples. The proposed method is evaluated on a public dataset, where it outperforms state-of-the-art approaches in challenging scenarios, and is validated in a real-world person-following experiment with a robotic platform in an environment with multiple distractors, occlusions, out-of-view situations and illumination changes. Despite these complexities, the robot consistently re-identified and followed the target individual. Runtime analysis demonstrates that the system operates reliably on embedded computing platforms with NVIDIA GPUs, making it both robust and resource-efficient for real-world deployment.**

I. Introduction

Robust Person Re-Identification (ReID) is a critical requirement for service robots to provide personalized assistance to a given target user, especially when executing guiding and following behaviors [1] in public environments. The Person Re-Identification problem has been tackled via different methodologies involving different sensing modalities, such as the use of radio-frequency devices held by the target user that provide a direct solution to the ReID problem [2], [3]. The main drawback of these approaches is the requirement of custom hardware and additional setup.

Vision-based approaches [4] are some of the most popular in the research community given that they rely merely on egocentric data coming from the camera sensors mounted on robotic platforms. Consequently, they require no additional setup in the environment, nor any device the user must carry. However, there are still many challenges in designing a system robust enough to operate in public spaces. Public spaces tend to have crowds that complicate the data association problem, causing issues such as the ID swap between tracks when a large amount of noisy detections and occlusions take place. This is especially problematic when trying to assist a specific target user. Also, people in the environment are not guaranteed to be wearing highly distinguishable features from one another. Thus, robust ReID becomes a challenge when there are people with similar appearances. Finally, people are not guaranteed to have a fixed appearance over time, since changes may come from illumination, background noise or changes in clothing (e.g. removing a jacket).

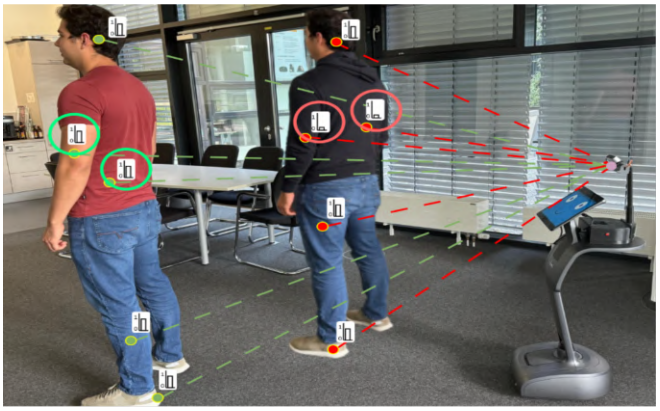


Fig. 1: Person ReID improves by learning which combination of body parts contribute the most to a robust global representation. Green and red circles indicate the key parts for identifying the target and the distractor, respectively.

This paper presents a person tracking and re-identification system targeted to service robots to be deployed in public spaces that can be used for HRI scenarios involving guiding and following behaviors. The proposed system uses Online Continual Learning (OCL) as presented in [5] to learn a robust representation through a One-Class Transformer network that uses the interaction between body-part features to re-identify a target person.

The following is a summary of the contributions made by this work:

- A continual learning framework for person re-identification using the One-Class Body-Part (OCBP) Transformer, which exploits interactions among body-part embeddings.
- Extended SORT [6] by integrating RGB-D data and introducing an exclusion-list mechanism to prevent the re-identification of known distractors during tracking.
- A memory management strategy that maintains sample diversity for online learning and provides pseudo-

This work is part of the Digitaler Bahnhof Minden (DiBaMi) Project funded by ERDF/JTF program NRW REGIONALE Ostwestfalen-Lippe - Networked Mobility and Digital Applications.

[1] Faculty of Technology at the University of Bielefeld, Inspiration 1, Bielefeld, Germany {jalemangallegos, swachsmu}@techfak.uni-bielefeld.de

Code repository: https://gitlab.ub.uni-bielefeld.de/dibami/ocbp_reid_ros

negative samples to increase the robustness of the transformer classifier and faster convergence.

The structure of the paper is as follows. Section II reviews prominent approaches and the state-of-the-art methods in people tracking and re-identification targeted towards service robot applications. Section III presents the proposed pipeline for person tracking and re-identification and detailed information about the different modules involved in the proposed solution. Section IV evaluates the performance of the proposed solution. Finally, Section V presents the conclusions of the work, discusses limitations and future work.

II. Related Work

A. Robo-centric People Tracking

Multiple Person Tracking is the problem of detecting, filtering and associating people over time. Approaches that focus on detecting and tracking people from a robot's point of view [7] face a different set of challenges than general Multiple Object Tracking (MOT) algorithms designed for static sensors. The main differences are the sensors' constrained Field of View (FoV) and the robot's motion in the environment.

Approaches vary depending on the sensing modality and the tracking architecture design. Examples such as [8] or [9] use an RGB-D camera to detect people through classical image and point cloud processing techniques and Bayesian tracking, achieving real-time performance. In [9], the system incorporates a learning methodology to train an AdaBoost classifier on hand-crafted features for each person being tracked, allowing association by considering both motion constraints and appearance similarity. Even though these approaches tackle the problem of long term tracking, appearance changes over time and identity preservation are not considered for re-identifying a specific target person after long occlusions or appearance changes.

Multimodal frameworks have been developed to benefit from detections of different sensor modalities with different arrangements at a given platform. Examples of such frameworks are presented in [10] and [11] where they loosely fuse each individual detection via aggregation techniques, Bayesian tracking and robust association algorithms. These approaches mainly benefit from the egocentric arrangements of the sensors on the robotic platforms to mitigate blind zones. A more recent approach for robo-centric people tracking for service robots is presented in [12], where Deep Learning based detections in both 2D and 3D, are used to track people with an omnidirectional camera and a 360° 3D LiDAR, an important aspect of this system is the coupling of the 2D and 3D information to ensure robust long-term target tracking and association.

B. Visual Person ReID for Service Robots

Person ReID is an additional module required when the identity of the people being tracked needs to be preserved in cases of occlusion and in-and-out-of-frame reappearance.

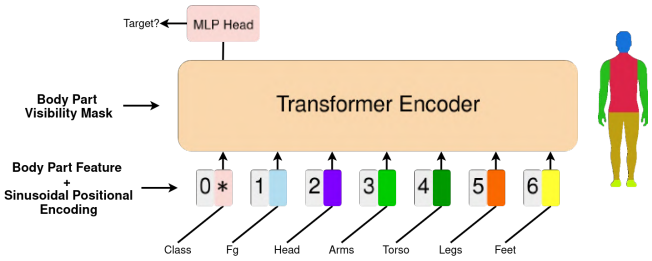


Fig. 2: One-Class Body-Part Transformer Classifier (OCBP-Transformer). This model takes six body-part embeddings plus a foreground embedding from a body-part feature extractor [15]. It follows the Vision Transformer [16] architecture with a *[CLS]* token and fixed sinusoidal positional encodings. One-Class classification takes place via an MLP on the *[CLS]* output embedding.

Tracking systems such as SORT [6] that initially did association purely considering motion constraints of detected bounding boxes, were later extended to integrate deep feature vectors that encode appearance information of the objects being tracked [13].

In a robotic person following application, the system developed in [14] provided a robust system that relied on a deep representation of the target person using a CNN as a feature extractor and an online boosting approach to train a classifier to take into account appearance changes and to re-identify a target person.

Another similar approach is described in [17] where different CNN architectures are trained online to classify if a given patch in an RGB image represents the target person. This approach trains the proposed classifiers online considering the target person patch as a positive sample and a nearby set of image patches (not necessarily with other people) as negative samples.

Some other alternatives such as [18] and [19] require an initialization/calibration step where several images of the target person are collected to have an initial generic representation. This accounts for multiple person viewpoints to improve the consistency before the robot starts any behavior. In each case, the metric used for ReID is cosine similarity and a statistical distance, respectively. In [18] multiple neural networks are used to detect the face and torso of a person and to extract features from each detection, still the initial representation is not updated over time. In contrast, in [19] a single global feature representation is updated via Damped Exponential Moving Average while also updating the decision threshold for re-identification.

Ye et al. [5] proposed a person tracking and re-identification pipeline that leverages a part-based feature extractor trained online via metric learning to enhance separation between the target and distractor features. The extractor, based on the method from [20], generates feature vectors for a fixed set of body parts, allowing the system to compare corresponding regions across individuals. During re-identification, particularly after long-term occlusion or out-of-frame disappearance, each body part's feature is processed by a per-part ridge regression classifier. Their outputs

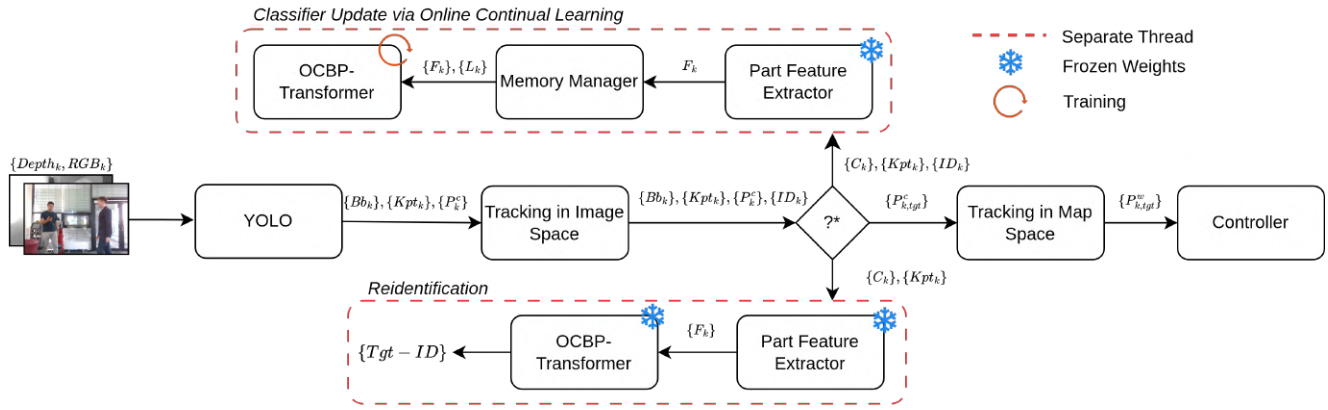


Fig. 3: Person Tracking and Re-Identification OCBP ReID Pipeline. At each time step k, RGB-D frames $\{RGB_k, Depth_k\}$ are processed using YOLOv11[23] to detect bounding boxes Bb_k and keypoints Kpt_k. The keypoints and the depth frame $Depth_k$ are then used to compute 3D positions of detected people with respect to the camera's optical frame. Bounding boxes and 3D poses are passed to the tracking module to assign track IDs. Once an ID is designated as the target, the algorithm checks in each frame whether the target is still detected. If present, the body-part feature extractor generates features F_k for each detection, which are stored in memory. The memory manager then selects a positive and negative (or pseudo-negative) sample pair to train the One-Class Body-Part Transformer Classifier online. If the target is not detected in frame k, body-part features are extracted for patches containing detected people, and the classifier determines whether any of them matches the target. A new ID is assigned to be the target if a candidate exceeds the confidence threshold $\delta_{\text{ReID}} = 0.8$. Finally, the target 3D pose is transformed to the map frame and tracked with a Linear Kalman Filter using the Constant Velocity (CV) model. The track is then sent to a controller that enables the robot to assist the user based on a specified behavior.

are averaged to determine whether a candidate matches the target. While the classifier is also trained online, it treats body parts independently, ignoring potential interactions between them and missing informative appearance relationships critical for robust re-identification.

Building on this, the present work adopts a part-based feature extractor [15], which is particularly effective in distinguishing individuals with similar overall appearance but differing in a few distinctive regions (e.g., trousers). The resulting body-part feature embeddings can be treated as a sequence of input tokens, making the transformer a natural choice for modeling the interactions between parts. Unlike pooling-based methods [21], transformers are well-suited for set- or sequence-like inputs [22], as they explicitly model relationships between input elements, an ability that is crucial when subtle inter-part cues are needed for accurate re-identification.

III. Methodology

A. Detection & Tracking

Figure 3 presents the pipeline of the proposed solution, which begins by detecting people in the current RGB frame at time k. This is accomplished by using the pretrained YOLOv11 model from Ultralytics [23], which provides bounding boxes Bb_k and 2D human poses Kpt_k outputs, the only class used for detection is the "*person*" class. Once both sets of detections are obtained, the $Depth_k$ frame is used, along with the keypoints to find the 3D position of each detected person with respect to the camera's optical frame ${P^c}_k$.

After detections are obtained, any tracking-by-detection method can assign and maintain consistent IDs over time. Examples of tracking systems used include: ByteTrack [24], SORT [6], and an extended SORT that leverages 3D detection data for improved ID association.

Purely 2D trackers are prone to ID swaps [25], where overlapping bounding boxes exchange IDs due to close proximity. In an OCL scenario where the labels used for training are extracted directly from the assumption of correct tracking, such errors are critical, as incorrect IDs can corrupt the training of a given model and cause incorrect re-identification.

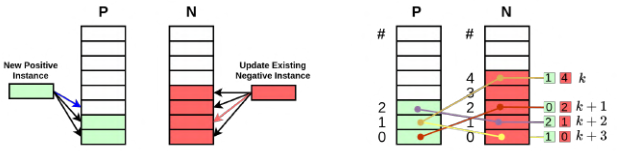
Some approaches [5] address this issue by re-identifying every bounding box before storing samples. To mitigate this, 3D data from the depth camera is used to enhance ID association. Given detections' positions in the camera frame ${P^c}_k$, SORT is extended by adding x^c and z^c coordinates and their velocities to the Kalman filter state as shown in Equation 1 (discarding y^c under the assumption that tracked people stand on the same plane as the robot).

$$x = [u\ v\ s\ r\ \dot{u}\ \dot{v}\ \dot{s}]^\mathsf{T} \rightarrow [u\ v\ s\ r\ x^c\ z^c\ \dot{u}\ \dot{v}\ \dot{s}\ \dot{x^c}\ \dot{z^c}]^\mathsf{T} \quad (1)$$

2D bounding boxes are still matched using the IoU [6] to form a cost matrix C_{IoU}, but this is complemented with a cost matrix C_{χ^2} derived from the Squared Mahalanobis distances between the (x^c, z^c) detections and the corresponding track state vectors. The final cost matrix is computed as in Equation 2, then passed to the Hungarian algorithm for detection-to-track assignment. This approach combines both spatial overlap in image space and Cartesian position constraints to resolve ambiguities in cases of overlapping tracks.

$$C = -C_{\text{IoU}} + C_{\chi^2} \quad (2)$$

Another important aspect is the implementation of an *exclusion list* mechanism to prevent re-identifying tracks



(a) Sample Insertion **(b)** Sample Retrieval

Fig. 4: Memory Manager Mechanism. a) Example of insertion when a positive sample is different from all the other existing samples (left) and when a new negative sample matches an existing sample (right). b) Four time steps of sample retrieval to enhance diversity, providing positive-negative pairs.

that have previously been confirmed not to be the target. This avoids unnecessary inference on known distractors when trying to re-identify the target and reduces the risk of incorrect re-identification when distractors overlap critical parts of the target. The main limitation of this approach is its vulnerability to ID swaps involving the target; therefore, the current tracking strategy for preventing ID swaps is preferred, as it mitigates this issue more effectively by incorporating both 2D and 3D information while promoting re-identification instead of loose association.

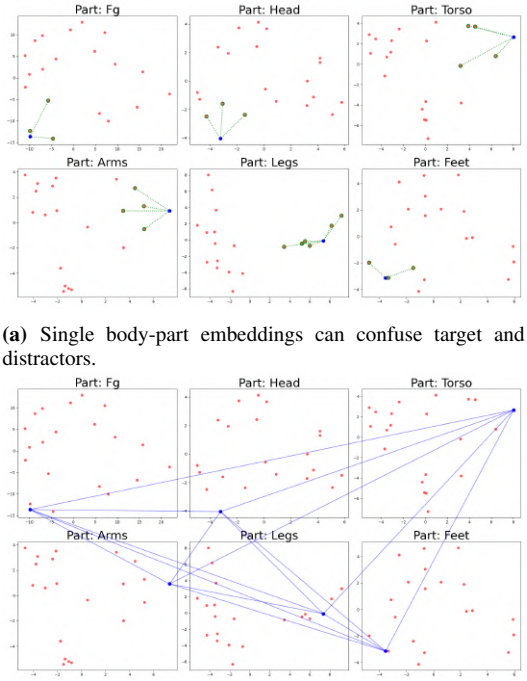
B. Memory Manager

During tracking, samples are collected to retrain the OCBP-Transformer (Figure 2). While the target ID is present in the current frame, body-part features and visibility masks are extracted using [15], from its corresponding bounding box patch and keypoints, then stored as a positive sample (label 1). All other detected people serve as negative samples (label 0).

1) Storing New Samples: To improve runtime efficiency and reduce memory usage, the batch size is constrained to one sample per image, alternating between positive and negative samples across frames to balance the samples when possible. At each time step k, only one sample is stored in the memory manager.

The memory manager maintains two equally sized data structures to store feature sets for positive and negative samples. When a new feature goes into the memory manager, it is first compared to all samples in the corresponding data structure using the part-based cosine distance [20] (Equation 3). Each body-part feature from set f^q is compared against all corresponding M stored embeddings f^m considering K parts and their visibilities v^m and v^q. If the distance is below a threshold δ_{sim} (indicating the sample is sufficiently different), the new sample is stored; otherwise, it replaces the most similar existing sample. The memory manager also tracks the number of times each sample has been updated (replaced), so when memory reaches capacity, it discards the identities with the smallest update count, Figure 4a exemplifies this insertion mechanism.

$$\forall\, m \in \{0, \ldots, M\} : \frac{\sum_{i=1}^{K} v_i^q v_i^m \operatorname{dist}_{\cos}(f_i^q, f_i^m)}{\sum_{i=1}^{K} v_i^q v_i^m} \qquad (3)$$



(a) Single body-part embeddings can confuse target and distractors.

(b) Interactions among body-part features contribute to a robust signature for person re-identification.

Fig. 5: a) Single body-part embeddings can lead to confusion, as the target may appear close to nearby distractors in the latent space. The edges illustrate the nearest negative features within an Euclidean distance of 0.5 (with normalized features) that could lead to confusion. **b)** Modeling interactions across body parts yields a more robust target representation.

2) Sample Retrieval: The memory manager returns either a positive–negative pair (if available) or a positive–pseudo-negative pair with probability β. If no negative samples have been collected, a positive–pseudo-negative pair is always used. As discussed in the following section and shown in Figure 7, this strategy accelerates convergence when training the transformer.

To tackle the problem of catastrophic forgetting when training the OCBP-Transformer online, the memory manager selects pairs so that no sample is repeated until all others have been served. This is achieved by randomly choosing samples while maintaining a counter for each, resetting the counters once all stored samples have been used. Figure 4b exemplifies the sample retrieval process.

C. One-Class Body-Part Transformer Classifier

In order to decide whether a detected individual corresponds to the *target person* or is instead a *distractor*, a **transformer encoder classifier network** is employed, as illustrated in Figure 2. This network takes as input the body-part embeddings extracted following the method of [15], consisting of six components: *Foreground*, *Head*, *Arms*, *Torso*, *Legs*, and *Feet*, each represented by a 512-dimensional feature vector. The classifier outputs a confidence score indicating whether a given sample corresponds to the target

person. Furthermore, the network incorporates the predicted visibility scores provided by the feature extractor, masking out the embeddings of body parts that are occluded, thereby ensuring robustness to partial visibility.

By leveraging a global representation that integrates all feature vectors and their mutual interactions, the model can effectively resolve local ambiguities caused by high similarity between individual body parts, as illustrated in Figure 5. In this way, the OCBP-Transformer learns to selectively attend to the most informative parts and their combinations, thereby constructing a more discriminative representation of the target person.

1) Online Learning: To handle appearance variations such as clothing or illumination changes, the OCBP-Transformer network is trained online using the Binary Cross Entropy Loss and the AdamW optimizer. Unlike [5], the feature extractor is not retrained; its weights remain fixed. Only the OCBP-Transformer Classifier is updated, using positive–negative or positive–pseudo-negative sample pairs with a fixed batch size of two to maintain balanced learning. Pseudo-negative samples are randomly drawn from a predefined distribution in the latent space, enabling the network to effectively learn the one-class classification (anomaly detection) task, as in [26], where a CNN is trained following this methodology. Training runs in a secondary thread and continues until the target ID is absent from the current frame, then the re-identification procedure is triggered.

2) Re-Identification: When the target ID is no longer present in the current frame, the re-identification procedure is triggered. In this step, features from all *non-exclusion-list* detections are considered as candidates. A high threshold δ_{ReID} is applied to ensure strict matching. The tracking system is essential at this point to prevent *exclusion-list* IDs from being swapped with a candidate that might be the target, which could otherwise lead to a deadlock where the target remains excluded.

IV. Evaluation

The proposed method for visual person tracking and re-identification is evaluated on the public sequences from [5], which contain targets and visually similar distractors with few distinctive cues. Ablation studies are conducted to assess the impact of the memory manager on OCL performance by systematically varying its parameters. The system is also deployed on a real robotic platform to assess the person-following behavior and report a runtime analysis to evaluate feasibility on resource-constrained devices.

A. Success Rate Evaluation in Public Dataset Sequences

The success-rate metric from [14, 17, 27, 5] is used for evaluating the performance in the public dataset sequences provided by [5]. For a sequence with N annotated frames, the success rate is defined as $SR = \frac{1}{N}\sum_{i=1}^{N} a_i \times 100\%$, where $a_i = 1$ indicates a correct match if the distance between the centers of the predicted and the annotated bounding boxes is below 50 pixels, otherwise $a_i = 0$.

TABLE I: Success Rate (%) of Person Tracking in Dataset [5]

Methods	Success Rate (%)			
	corridor1	corridor2	lab-corridor	room
Zhong's Method	63.8	66.8	75.8	44.7
SiamRPN++	44.8	55.9	46.1	42.6
STARK	44.3	83.8	73.1	65.8
SORT + RPF-ReID	67.3	37.9	31.1	82.4
OC-SORT + RPF-ReID	67.3	37.9	31.1	82.4
ByteTrack + RPF-ReID	69.1	20.2	54.2	82.4
ByteTrack + RPF-ReID + OCL	93.5	94.9	**96.0**	96.8
SORT+OCBP ReID	91.02	95.71	89.71	**98.88**
ByteTrack+OCBP ReID	93.65	96.12	93.50	98.51
Depth-Aware SORT+OCBP ReID	**96.20**	**97.02**	94.50	97.64

*This table extends Table I from [5]. The results from [5] are reported as published and were not reproduced in this work. The additional rows correspond to the proposed system, and are highlighted in gray. Bold values indicate the best performance.

Since the sequences in [5] do not provide depth data, DepthPro [28] was employed to generate metric depth images. These depth maps enabled a fair evaluation of the proposed depth-aware tracker (Sec. III) along with SORT and ByteTrack in the OCBP ReID pipeline. The proposed approach outperforms ByteTrack + RPF-ReID+OCL [5] on three of the four sequences with strong target–distractor similarity, frequent occlusions, and in/out-of-frame events. In *corridor1*, *corridor2* and *room*, the improvement exceeds 2% over the previous approach. The *lab-corridor* sequence is more challenging due to re-entries, appearance changes, and severe illumination shifts. Nevertheless, the target is consistently recovered and tracked until the end of the sequence in all cases.

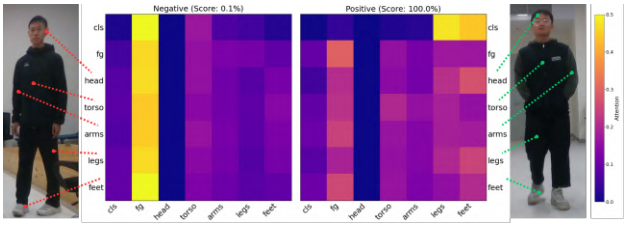


Fig. 6: Attention visualization of OCBP-Transformer for a target person (right) and a distractor person (left) from the *room* sequence, highlighting the attention to the most distinctive body part, in this case being the *legs* and *feet*.

1) The Role of the One-Class Body-Part Transformer Classifier: The OCBP-Transformer aggregates information across body-part embeddings to decide whether a detection matches the target. As illustrated in Figure 6, the model focuses on the most discriminative parts for the target (e.g., legs and feet), while assigning higher attention from the $[CLS]$ token to the foreground and specific parts when inferring a distractor. Since the classifier is trained online, attention adapts to whichever features best separate the target from nearby candidates in the current scene.

2) Use of Pseudo-Negative Samples: Following [26], pseudo-negative samples are drawn from a Gaussian distribution to stabilize the one-class (anomaly detection) training when real negatives are scarce. This is especially useful early in a run when only positive samples are available and

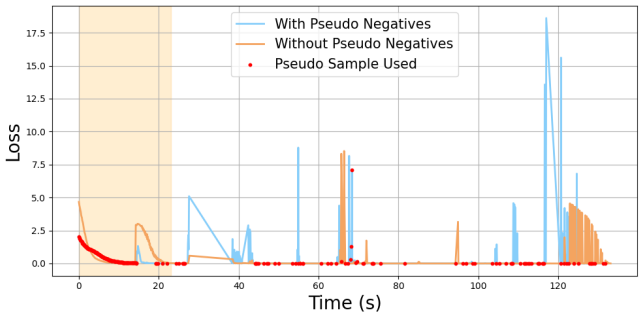


Fig. 7: Effect of pseudo-negative samples in the One-Class Body-Part Transformer Classifier during Online Learning over the *corridor1* sequence.

distractors have not yet appeared. As shown in Figure 7, during roughly the first 20 s (highlighted), pseudo-negatives balance the training batches and speed up convergence; thereafter, they are used stochastically ($\beta = 0.1$) to maintain diversity and avoid overfitting.

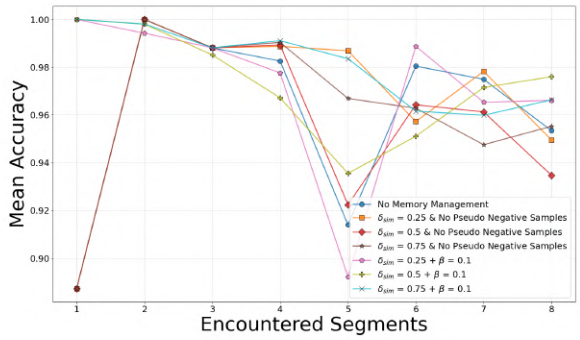
B. Ablation Studies on the Memory Management System

To evaluate the impact of the proposed memory management system, an ablation study was conducted by varying key parameters, particularly the similarity threshold δ_{sim}, which regulates sample diversity within the memory. A low δ_{sim} value tends to replace more samples, resulting in a compact memory with limited diversity. While this reduces redundancy, it restricts the system's ability to support generalization during replay. Conversely, a high δ_{sim} value treats most incoming samples as distinct, yielding a larger and more diverse memory. This balance directly influences the model's generalization capacity.

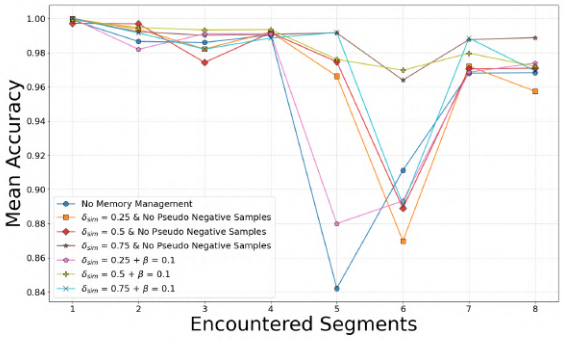
Figure 8 reports the average classification accuracy across multiple sequences from [5], following the methodology in [29, 30]. Each sequence was divided into eight segments, and the classifier's accuracy was computed iteratively by training on the current segment and evaluating on all previous ones. Re-Identification was considered correct when $\delta_{\text{ReID}} \geq 0.8$ and the ground truth annotated bounding boxes were used to discard the impact of the tracker.

Even without memory management, the OCBP-Transformer Classifier maintained an accuracy above 80% across all sequences, demonstrating strong retention over time. However, incorporating the proposed memory management system further improved performance, often surpassing 90%, particularly when $\delta_{sim} \geq 0.5$ and pseudo-negative samples were included.

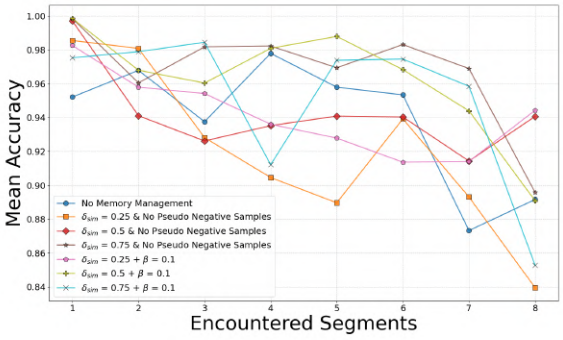
The role of pseudo-negatives is especially evident in scenarios such as the *corridor1* sequence, where initially only positive samples were available until a distractor appeared with strong appearance similarity. As shown in Figure 8a, the presence of pseudo-negatives substantially improved early-segment performance by enabling faster adaptation with an accuracy greater than 98%. Once real negatives are available, pseudo-negatives are sampled with probability β. This consistently improves learning speed and generalization



(a) Corridor 1



(b) Corridor 2



(c) Lab Corridor

Fig. 8: Average accuracy over eight evaluation segments for Online Continual Learning. Results are shown for: (i) no memory management, (ii) memory management with three variants of δ_{sim} controlling sample diversity, and (iii) experiments with pseudo-negative samples (β), as described in Sec. III, to assess their impact on generalization for Re-Identification purposes.

across all sequences. Overall, the results highlight that combining a diversity-promoting δ_{sim} with pseudo-negative sampling enables the OCBP-Transformer to adapt more rapidly and achieve robust generalization, even in challenging re-identification scenarios.

C. Runtime Analysis

The deployment of person tracking and re-identification systems is often constrained by their high computational

requirements. Since the integration of robotic systems into public environments demands energy-efficient solutions that can scale, the runtime performance of the proposed ReID pipeline is evaluated on the three NVIDIA Jetson Orin platforms available at the time of publication. Table II summarizes the runtime performance, measured in *milliseconds*, of the main system components illustrated in Figure 3, with particular emphasis on those modules that rely on GPU acceleration for practical deployment. All test setups were configured with *Jetson clocks* enabled and maximum power mode activated.

TABLE II: Performance comparison across Jetson Orin platforms (mean / min / max / std in ms).

Platform	Detection	Tracking	ReID*	Train*
Nano	55.77	10.40	653.84	692.06
	24.37	1.40	589.20	589.45
	219.23	69.32	789.70	822.98
	34.74	12.82	91.69	49.44
NX	39.31	1.93	337.50	351.49
	17.43	1.04	335.24	333.75
	122.92	11.53	339.77	366.61
	26.54	1.48	2.27	7.34
AGX	30.04	1.66	282.11	273.01
	15.55	0.93	261.74	252.77
	88.95	10.78	302.31	297.51
	14.84	1.10	16.61	9.50

*ReID and Training are executed in separate threads.

The system operates on RGB-D frames at 5 Hz. As shown in Table II, the average tracking and detection time (highlighted in gray) remains within this limit across all evaluated embedded platforms, ensuring reliable deployment on low-power devices. The detection module (YOLOv11) is the only component optimized with TensorRT and compiled using FP16 quantization. The reported runtime, therefore, includes not only model inference but also keypoint processing, pose extraction, and data preparation required by other modules.

Although the OCBP-Transformer online training time exceeds the sensor frequency, it runs in a separate thread and does not interfere with the main pipeline. Similarly, while the ReID module may take longer than the per-frame update, it provides a new target ID as soon as a person is re-identified, without disrupting the ongoing tracking-by-detection process.

D. Robustness in Person Following Task

The performance of the re-identification system is further evaluated in a person-following application, where a target individual follows a predefined trajectory and is occasionally occluded by distractors. The robot's objective is to maintain a desired distance (1.4 m from the robot's center) from the target, using the controller proposed in [31]. For this evaluation, a customized Agilex Ranger Mini v3 platform, adapted to assist users in carrying luggage, is employed. The robot is equipped with two Intel RealSense D456 RGB-D cameras, an NVIDIA Jetson Orin NX, and a Zotac CI669

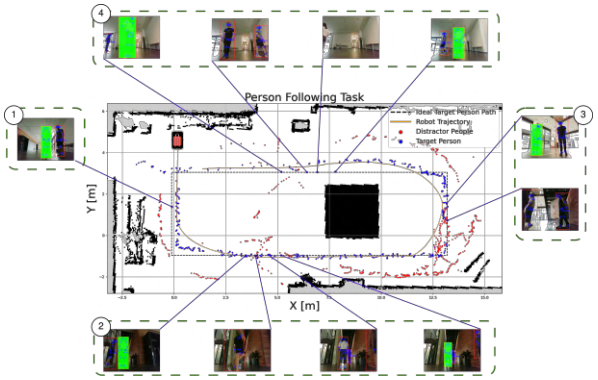


Fig. 9: Person-following robot experiment with OCBP-Transformer + depth-aware SORT. (1) The target person is detected. (2) A distractor appears, causing an occlusion that requires re-identification. (3) Another re-identification event occurs. (4) The target person leaves the field of view while distractors remain, and is successfully re-identified upon reappearance.

onboard computer powered by an *Intel® Core™ i7-1355U CPU*. The entire system runs on ROS 2 Humble.

As shown in Figure 9, the proposed re-identification system demonstrates robust performance under occlusions, out-of-frame re-entries, and the presence of multiple distractors. It operates reliably on low-power embedded GPU platforms and is resilient to illumination changes, enabling the robot to consistently track the target until the end of the trajectory.

V. Conclusions

This work presented a novel person re-identification system for human–robot interaction, built on the proposed One-Class Body-Part (OCBP) Transformer Classifier. Trained online, the model leverages body-part feature interactions to construct a robust representation of the target person. Its continual learning process is supported by an enhanced tracking strategy, which excludes distractor tracks, and a memory management mechanism that mitigates catastrophic forgetting.

Ablation studies were conducted to evaluate the performance of different configurations of the memory management system which especially highlight the benefit of the replay strategy and the need for pseudo-negative samples to improve robustness and encourage faster convergence, having a special positive impact in cases where initially only the target person is present and when distractors suddenly appear with high appearance similarity. The proposed system outperforms state-of-the-art methods on a public benchmark, achieving high success rates under challenging conditions, including strong visual similarity between the target and distractors, frequent occlusions, and temporary loss of the target from the field of view. Crucially, it operates efficiently on embedded NVIDIA Jetson Orin platforms, enabling deployment in real-world robotic systems. Validation on a person-following task confirmed its robustness, while the approach remains applicable to broader use cases such as guiding behaviors and engagement detection.

While the integrated tracker improves robustness, further reliability could be achieved through bidirectional human–robot communication, benefiting from the interaction itself [32] and ensuring correct target selection without the need for additional sensing modalities.

Future research may explore transformer-based few-shot learning approaches (e.g., [33]), which can generalize from limited samples and progressively improve as more data becomes available, without requiring continuous retraining.

REFERENCES

[1] Andrea Eirale, Mauro Martini, and Marcello Chiaberge. "Human Following and Guidance by Autonomous Mobile Robots: A Comprehensive Review". In: *IEEE Access* 13 (2025), pp. 42214–42253.

[2] Carmen Scheidemann et al. "Obstacle-avoidant leader following with a quadruped robot". In: *arXiv preprint arXiv:2410.00572* (2024).

[3] Dingzhi Zhang et al. "A Hybrid Human Tracking System using UWB Sensors and Monocular Visual Data Fusion for Human Following Robots". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2024, pp. 10878–10883.

[4] Md Jahidul Islam, Jungseok Hong, and Junaed Sattar. "Person-following by autonomous robots: A categorical overview". In: *The International Journal of Robotics Research* 38.14 (2019), pp. 1581–1618.

[5] Hanjing Ye et al. "Person Re-Identification for Robot Person Following With Online Continual Learning". In: *IEEE Robotics and Automation Letters* 9.11 (2024), pp. 9151–9158.

[6] Alex Bewley et al. "Simple online and realtime tracking". In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468.

[7] Viktor Schmuck and Oya Celiktutan. "RICA: Robocentric Indoor Crowd Analysis Dataset". In: *UKRAS20 Conference: "Robots into the real world" Proceedings* (2020), pp. 63–65.

[8] Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. "Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 5636–5643.

[9] Matteo Munaro and Emanuele Menegatti. "Fast RGB-D people tracking for service robots". In: *Autonomous Robots* 37.3 (Oct. 2014), pp. 227–242. ISSN: 1573-7527.

[10] Timm Linder et al. "On multi-modal people tracking from mobile platforms in very crowded and dynamic environments". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 2016, pp. 5512–5519.

[11] Tim Wengefeld et al. "A Multi Modal People Tracker for Real Time Human Robot Interaction". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2019, pp. 1–8.

[12] A. Shenoi et al. "JRMOT: A Real-Time 3D Multi-Object Tracker and a New Large-Scale Dataset". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 10335–10342.

[13] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. "Simple Online and Realtime Tracking with a Deep Association Metric". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 3645–3649.

[14] Kenji Koide, Jun Miura, and Emanuele Menegatti. "Monocular person tracking and identification with on-line deep feature selection for person following robots". In: *Robotics and Autonomous Systems* 124 (2020), p. 103348. ISSN: 0921-8890.

[15] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. "Keypoint Promptable Re-Identification". In: *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Nov. 2024, pp. 216–233. ISBN: 9783031729867.

[16] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR* (2021).

[17] Bao Xin Chen, Raghavender Sahdev, and John K. Tsotsos. "Integrating Stereo Vision with a CNN Tracker for a Person-Following Robot". In: *Computer Vision Systems*. Ed. by Ming Liu, Haoyao Chen, and Markus Vincze. Cham: Springer International Publishing, 2017, pp. 300–313. ISBN: 978-3-319-68345-4.

[18] Mario Srouji, Jian Zhang, and Hugues Thomas. "Human Following in Mobile Platforms with Person Re-Identification". In: *CASE*. 2024.

[19] Federico Rollo et al. "Continuous Adaptation in Person Re-identification for Robotic Assistance". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 2024, pp. 425–431.

[20] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. "Body Part-Based Representation Learning for Occluded Person Re-Identification". In: *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2023, pp. 1613–1623.

[21] Manzil Zaheer et al. "Deep Sets". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3391–3401.

[22] Juho Lee et al. "Set transformer". In: *International Conference on Machine Learning*. Vol. 4. 8. 2019.

[23] Glenn Jocher and Jing Qiu. *Ultralytics YOLO11*. Version 11.0.0. 2024.

[24] Yifu Zhang et al. "ByteTrack: Multi-object Tracking by Associating Every Detection Box". In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Tel Aviv, Israel: Springer-Verlag, 2022, pp. 1–21. ISBN: 978-3-031-20046-5.

[25] Keni Bernardin and Rainer Stiefelhagen. "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics". In: *EURASIP Journal on Image and Video Processing* 2008.1 (May 2008), p. 246309. ISSN: 1687-5281.

[26] Poojan Oza and Vishal M. Patel. "One-Class Convolutional Neural Network". In: *IEEE Signal Processing Letters* 26.2 (Feb. 2019), pp. 277–281. ISSN: 1558-2361.

[27] Hanjing Ye et al. "Robot Person Following Under Partial Occlusion". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)* (2023), pp. 7591–7597.

[28] Aleksei Bochkovskii et al. "Depth Pro: Sharp Monocular Metric Depth in Less Than a Second". In: *International Conference on Learning Representations*. 2025.

[29] David Lopez-Paz and Marc'Aurelio Ranzato. "Gradient episodic memory for continual learning". In: *Advances in neural information processing systems* 30 (2017).

[30] Zheda Mai et al. "Online continual learning in image classification: An empirical survey". In: *Neurocomputing* 469 (2022), pp. 28–51. ISSN: 0925-2312.

[31] Julio Montesdeoca et al. "Person-Following Controller with Socially Acceptable Robot Motion". In: *Robotics and Autonomous Systems* 153 (2022), p. 104075. ISSN: 0921-8890.

[32] Federico Rollo et al. "FollowMe: a Robust Person Following Framework Based on Visual Re-Identification and Gestures". In: *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. 2023, pp. 84–89.

[33] Adrian Bulat et al. "FS-DETR: Few-Shot DEtection TRansformer with Prompting and without Re-Training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 11793–11802.