electronics

MDPI

Article

Demographic-Aware Multi-Object Tracking for Retail Environments via Temporal Consistency and Joint Association

Iason-Ioannis Panagos [1,*], Angelos P. Giotis [1], Marina E. Plissiti [1], Vasiliki Stamati [1], George Gartzonikas [1], Michalis Vrigkas [2] and Christophoros Nikou [1]

[1] Department of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece; g.gkartzonikas@uoi.gr (G.G.); cnikou@uoi.gr (C.N.)
[2] Department of Communication and Digital Media, University of Western Macedonia, 52100 Kastoria, Greece
* Correspondence: i.panagos@uoi.gr

Abstract

Reliable identity preservation is essential in retail multi-object tracking because an identity switch may assign dwell time, shelf interactions, or demographic statistics to the wrong consumer trajectory. This challenge is amplified in crowded indoor scenes, where occlusions and appearance ambiguity weaken conventional association cues. This work presents a demographic-aware multi-object tracking framework for retail environments that exploits spatial, motion-based, appearance-based, and demographic cues within a modular tracking-by-detection pipeline. The proposed approach integrates IoU-based spatial association, LSTM-based motion forecasting, ReID appearance embeddings, and apparent demographic information derived from age-group and gender predictions into a common association cost. The core assumption is that demographic attributes should remain consistent across neighboring frames for the same tracked individual; therefore, demographic agreement can serve as a weak semantic cue during detection-to-tracklet association. Unlike conventional pipelines that treat multi-object tracking and demographic estimation as independent stages, the proposed method reuses the outputs of an existing Inception-based apparent demographic classifier to support identity preservation, without introducing a new demographic estimation model or requiring end-to-end retraining. The framework is evaluated on RGB retail tracking sequences from the Consumers dataset under both raw and privacy-aware anonymized settings. The proposed method achieves 84.8% MOTA and 87.8% IDF1 on raw data and 80.0% MOTA and 85.3% IDF1 under anonymization, improving the strongest evaluated baseline by 0.4 and 0.6 percentage points and by 1.0 and 1.3 percentage points, respectively, while preserving the same low ID-switch counts. Although incremental in absolute terms, these gains are consistent across both evaluation settings and are obtained over a strong baseline in challenging indoor retail sequences. The results indicate that even a lightweight demographic-consistency cue can provide measurable complementary information for improving consumer trajectory stability.

Keywords: multi-object tracking; demographic estimation; ReID; retail analytics; tracking-by-detection

Academic Editor: Kefeng Ji

Received: 31 May 2026
Revised: 3 July 2026
Accepted: 11 July 2026
Published: 14 July 2026

1. Introduction

1.1. Problem Context and Motivation

The automation of consumer behavior understanding in physical retail environments remains an important challenge for computer vision and retail analytics. Multi-object

tracking (MOT) forms the algorithmic foundation of this process, since per-frame detections must be associated into temporally coherent trajectories before higher-level indicators can be computed. In retail scenarios, such trajectories enable the estimation of dwell time, zone affinity, shopper–shelf interaction, and inter-consumer movement patterns that can support store-layout analysis and product-placement strategies [1,2].

In this setting, overall tracking accuracy is not the only relevant objective. For downstream retail analytics, preserving stable identities over time is particularly important, because identity switches may propagate directly to estimates of dwell time, shelf/product association, and demographic group-level statistics. A fragmented or inconsistent tracklet may therefore affect not only standard MOT indices, such as MOTA and MOTP, but also the reliability of subsequent consumer-behavior indicators. Within this framework, reducing identity switches and improving tracklet continuity becomes essential for transforming visual observations into meaningful retail analytics.

Indoor retail scenes introduce domain-specific challenges that are not always fully captured by standard pedestrian tracking benchmarks. Consumers often move irregularly, stop near shelves, turn abruptly, interact with products, or become partially occluded by other consumers and store structures. Reflective surfaces, narrow aisles, low-resolution body crops, and visually similar clothing further increase the ambiguity of frame-to-frame association. The authors in [3] addressed this domain gap by introducing the *Consumers* and Body-Image Dataset (*BID*), captured in live retail stores, together with a modular pipeline combining YOLOX-based detection, LSTM-augmented motion forecasting, and apparent demographic classification at real-time speeds.

Raw movement trajectories can be further enriched with apparent demographic attributes, such as age group and perceived gender, to support demographic-level retail analysis. This estimation layer operates on full-body bounding boxes without relying on facial imagery, both because retail camera viewpoints frequently fail to capture the face and because privacy constraints and data-protection requirements limit unconstrained facial data processing in public commercial spaces. The authors in [3] further reported that an Inception-based multi-attribute classifier can retain competitive performance on anonymized body crops, suggesting that silhouette, posture, and clothing can provide useful demographic cues for practical consumer segmentation without exposing identifying facial information.

1.2. Challenges and Limitations of Existing Methods

1.2.1. Multi-Object Tracking

Multi-object tracking (MOT) remains a challenging computer vision problem because it requires the continuous association of detections across time under changing visual and scene conditions. Although recent deep learning-based tracking frameworks have significantly improved tracking accuracy and robustness, maintaining reliable trajectory continuity and consistent identity assignment in real-world scenarios remains difficult [4,5]. These challenges are particularly relevant in indoor retail environments, where consumers are often captured under low resolution, varying illumination, cluttered backgrounds, reflective surfaces, and significant viewpoint changes. In such settings, MOT systems do not only aim to maximize standard tracking accuracy; they also provide the temporal structure required by downstream analytics, including dwell-time estimation, zone occupancy, shelf/product association, and demographic-level behavioral analysis. Consequently, ID switches and fragmented trajectories may propagate errors to subsequent inference stages, reducing the reliability of the final consumer analytics.

One of the most significant challenges in MOT is occlusion, which occurs when tracked individuals become partially or fully hidden by other pedestrians, products, shelves,

or scene structures. Occlusions are especially common in crowded or narrow indoor spaces and often interrupt trajectory continuity, leading to fragmented tracks or incorrect identity reassignment after target reappearance. Traditional tracking approaches based mainly on spatial proximity or bounding-box overlap frequently struggle under prolonged occlusions due to the loss of visual continuity between consecutive detections [6]. Similarly, appearance-based methods may fail when only limited body regions remain visible or when the visual evidence is degraded by blur, low resolution, or strong viewpoint changes. To improve robustness, recent approaches increasingly incorporate motion prediction, temporal reasoning, and attention-based mechanisms capable of exploiting longer-range dependencies [7–9].

Another major challenge is identity switching, where the identity assigned to a tracked individual changes incorrectly during tracking. This typically occurs due to occlusions, fast motion, missed detections, or visually similar individuals. ID switches directly affect trajectory consistency and are particularly harmful in retail-oriented MOT, since a single identity mismatch may assign dwell time, shelf interactions, or demographic statistics to the wrong consumer trajectory. Modern MOT frameworks attempt to reduce ID switches through the integration of motion models, appearance embeddings, and re-identification (ReID) networks [10–12]. Nevertheless, maintaining stable identities over longer temporal intervals remains an open research problem, especially in crowded indoor scenes where multiple consumers may exhibit similar clothing, posture, or body shape.

Appearance ambiguity is therefore a fundamental limitation of many tracking-by-detection approaches. In pedestrian and consumer tracking scenarios, ambiguity often emerges when different individuals share visually similar characteristics, such as clothing color, body proportions, or movement patterns. Environmental factors, including illumination changes, motion blur, and low-resolution video, further reduce the discriminative capability of visual features. In privacy-sensitive retail settings where reliable face information is unavailable or intentionally avoided, systems rely primarily on body-level features, increasing the risk of incorrect matches and ID switches. To alleviate these issues, recent approaches combine appearance descriptors with temporal, spatial, and geometric cues [13], while transformer-based architectures have shown improved robustness by capturing richer spatio-temporal dependencies [9,14].

Reliable data association is therefore central to MOT, as it links detections across frames to form trajectories. Many modern trackers rely on either intersection-over-union (IoU)-based matching or appearance-based ReID embeddings. IoU-based association assumes spatial continuity between consecutive detections and associates objects based on bounding-box overlap. While computationally efficient, this strategy becomes unreliable under rapid motion, occlusions, or missed detections [6]. ReID-based association relies on deep appearance embeddings to distinguish between targets, but its performance may degrade under viewpoint changes, illumination variation, partial visibility, or low resolution [10]. Hybrid approaches that combine IoU matching, motion estimation, and appearance embeddings provide improved robustness in complex scenarios [3,12,15]. However, these strategies typically rely on geometric and appearance-based evidence only. In contrast, the present work investigates whether apparent demographic predictions can provide an additional weak semantic consistency cue during association, complementing spatial, motion, and appearance information in challenging retail tracking scenarios.

1.2.2. Demographic Predictions

Frame-level demographic predictions may exhibit temporal inconsistencies when applied independently to each detected body crop. The same consumer may receive different age-group or gender predictions across neighboring frames due to pose variation,

partial occlusion, motion blur, low image resolution, or proximity to a classifier decision boundary. Although such fluctuations may have limited impact when predictions are evaluated independently at the frame level, they become more problematic when demographic information is aggregated over tracklets for downstream retail analytics.

Most existing pipelines treat tracking and demographic estimation as separate stages: first, detections are associated into trajectories, and then demographic predictions are assigned to the corresponding body crops. Under this formulation, the demographic classifier has no direct influence on the association decision, even though demographic attributes should remain stable for the same individual across neighboring frames. Conversely, candidate detections with compatible demographic predictions can provide weak semantic evidence that they may belong to the same identity, especially when geometric and appearance-based cues are ambiguous.

Instead of using demographic predictions only as post-processing information over completed trajectories, this work exploits them at an earlier stage, during frame-to-tracklet association. Apparent age-group and gender predictions are used as semantic consistency cues within the association cost, complementing IoU-based spatial matching, motion prediction, and appearance-based ReID embeddings.

1.3. Proposed Approach and Main Contributions

In this work, we propose a demographic-aware multi-object tracking framework for retail environments that integrates demographic information directly into the association stage of a tracking-by-detection pipeline. The proposed method combines spatial information derived from bounding-box overlap, motion cues provided by the existing LSTM-based tracker, appearance-based similarity extracted from a ReID module, and demographic consistency cues obtained from age-group and gender predictions.

The core assumption of the proposed approach is that demographic attributes of the same individual remain consistent across neighboring frames. Based on this observation, demographic predictions are incorporated into the association cost matrix together with IoU- and embedding-based distances in order to improve identity preservation and reduce mismatches during tracklet assignment. As a matter of fact, the proposed work does not introduce a new demographic classifier. On the contrary, it investigates how the outputs of an existing Inception-v3-based apparent demographic estimation model can be reused as semantic association cues within a MOT pipeline. The proposed formulation is modular and can be integrated into existing tracking-by-detection pipelines without retraining the detector or the demographic classifier.

The main contributions of this work can be summarized as follows:

- We introduce a demographic-aware association strategy for multi-object tracking in retail environments, where apparent age-group and gender predictions are incorporated into the detection-to-tracklet matching cost.
- We formulate a heterogeneous association cost that combines spatial overlap, ReID-based appearance similarity, and demographic consistency, including an ordinal age-group distance term.
- We evaluate the proposed association strategy on retail-oriented RGB tracking sequences under privacy-aware conditions and compare it against established MOT baselines.
- We show that incorporating demographic consistency can improve identity preservation and tracklet stability in challenging indoor retail scenes, supporting downstream consumer analytics where stable identities are critical.

1.4. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work on multi-object tracking, ReID-based association, demographic estimation, and attribute-aware tracking. Section 3 presents the proposed methodology, including the problem formulation, the underlying tracking-by-detection framework, the demographic-aware association cost, and the Hungarian-based tracklet update strategy. Section 4 describes the experimental setup, including the dataset, evaluation metrics, and implementation details. Section 5 reports the quantitative results and comparative analysis, focusing on tracking performance, identity preservation, and runtime behavior. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related Work

2.1. Multi-Object Tracking in Retail Environments

Multi-object tracking (MOT) in retail environments is commonly formulated under the tracking-by-detection (TBD) paradigm, where detections are linked across frames to form temporally coherent trajectories. This formulation is compatible with modern deep learning-based object detectors and supports the extraction of higher-level retail analytics, such as customer trajectories, dwell time, zone occupancy, and shopper–shelf interactions. However, retail environments introduce domain-specific challenges, including frequent occlusions, narrow aisles, reflective surfaces, irregular customer motion, and visually similar consumers.

Early work in retail tracking introduced trajectory-based customer behavior analysis using Bayesian tracking methods in store environments [16]. Later, a real-time tracking framework for retail traffic-flow analysis based on periodic detections and lightweight data association was proposed in [17].

Computer vision-based smart shelf systems further associate customers with product interactions using tracking and filtering strategies [18]. More recently, deep learning-based systems have been introduced for consumer behavior analysis in real retail environments [19,20]. In a scenario involving multiple cameras, Mendes et al. [21] incorporated trajectory-based person re-identification to maintain customer identities across non-overlapping cameras. Beyond camera-based tracking, several retail analytics systems have explored alternative sensing modalities. RFID-based tracking data and pattern mining have been used to analyze shopping paths and their relationship to purchasing behavior [22–24].

Multi-sensor approaches, including RGB-D and RGB-thermal configurations, have also been explored in person-tracking and retail-related scenarios, mainly to improve robustness under illumination changes, occlusions, and partial visibility. For example, RGB-D vision and radio beacon sensor fusion have been investigated for customer profiling and indoor movement analysis [25], while BLE beacon-based localization and smart shopping-cart systems have been proposed for customer positioning, heat-map analysis, and personalized in-store advertising [26,27]. A retail-oriented multi-modal tracking-by-detection framework combining RGB and thermal imagery for consumer tracking and demographic analysis was presented in [28], while RGB-thermal fusion methods have been proposed to improve detection robustness under challenging visual conditions [29,30]. Although these works demonstrate the potential of multi-modal sensing, the present study focuses on RGB retail tracking sequences and investigates how demographic predictions can be exploited as an additional association cue.

Beyond retail-oriented MOT, recent deep learning approaches from adjacent detection and tracking domains further illustrate the benefits of context-aware and discriminative representation learning under challenging sensing conditions. Wang et al. [31] introduced

a global feature-injected blind-spot network for hyperspectral anomaly detection, using enlarged receptive fields and channel-attention-based global context modeling to improve the separation of anomalous targets from complex backgrounds. Teng et al. [32] proposed a positive-negative prompt mining network for UAV-based multispectral object tracking, in which target-relevant information and background-interference cues are jointly modeled to strengthen the learned tracking representation. Although these methods address different tasks and sensing modalities and are therefore not directly comparable with RGB retail MOT, they demonstrate the broader value of global-context modeling and explicit target-background discrimination in deep detection and tracking systems.

Privacy-aware tracking and anonymization methods have also emerged to address data protection concerns in video analytics that aim to reduce personally identifiable information while preserving downstream visual utility. Existing approaches include reversible image transformation [33,34], head inpainting [35], privacy-enhanced video streaming [34], face anonymization [36,37] and configurable video deanonymization [38] as representative examples. In the present work, privacy awareness is addressed at the dataset and evaluation level through anonymized body crops and face-blurred consumer sequences, following the retail benchmark introduced in [3].

2.2. Appearance-Based Association

Recent MOT approaches have increasingly explored stronger association mechanisms, including transformer- and attention-based association models [9] and ReID-assisted multi-stage matching strategies [12]. Joint detection-and-embedding architectures, such as JDE [39] and FairMOT [40], learn detection and ReID features within a single network, reducing the latency overhead of running a separate ReID model. Furthermore, Chan et al. [41] proposed a multi-object tracking framework that combines hybrid attention mechanisms and embedding association to improve feature discrimination and reduce identity switches by jointly exploiting appearance and geometric information during object association.

Apart from architectural unification, works such as DeepSORT [10] and StrongSORT [11] have adapted appearance embeddings to the requirements of MOT, which differ from standard person ReID benchmarks. More recently, Wang et al. [42] proposed a method to improve robustness under occlusion, while a transformer-based ReID extension for stronger representations under occlusions and long temporal gaps is explored in [43]. Cross-resolution ReID approaches [44] further attempt to address degraded visual quality, although such methods are still rarely integrated into practical MOT pipelines. Despite these advances, appearance-based association remains challenging under severe occlusions, similar clothing, partial visibility, and low-resolution detections.

2.3. Apparent Age and Gender Estimation

Apparent demographic estimation aims to infer visually perceived attributes, such as age group and gender, from image data. Most existing approaches rely primarily on facial information, since facial texture, geometry, and appearance provide strong cues for age and gender estimation [45–47]. However, in retail environments, facial information is often unavailable, unreliable, or intentionally anonymized due to camera viewpoint, low resolution, occlusions, and privacy constraints. As a result, body-based apparent demographic estimation becomes a more suitable setting for privacy-aware retail analytics.

Body-based demographic estimation relies on coarser visual cues, including silhouette, posture, clothing, body shape, and gait-related information [48–50]. Although these cues are less discriminative than facial information, previous work has shown that they can support practical age-group and gender prediction when facial information is not available.

In particular, Panagos et al. [3] introduced the Body-Image Dataset (*BID*) and evaluated an Inception-based multi-attribute classifier on full-body consumer crops under anonymized conditions, while Yuan et al. [51] proposed the Body Age (BAG) dataset for age estimation using body-based images.

In the present work, we do not introduce a new age or gender estimation model. Instead, we reuse the outputs of the existing Inception-based demographic classifier as semantic cues for data association. This shifts the role of demographic estimation from a purely downstream analytics module to an auxiliary source of identity-consistency information during tracking.

2.4. Attribute-Aware Association for Tracking

Semantic pedestrian attributes, including gender, age, clothing color, body shape, and hairstyle, provide high-level identity-related information that may complement low-level appearance embeddings.

Published work in pedestrian attribute recognition and soft biometrics showed that such descriptors can support person re-identification when facial information is unavailable or when visual conditions reduce the reliability of texture-based descriptors [52,53].

Attribute-assisted ReID methods further demonstrated that combining attribute-level information with conventional appearance descriptors can improve identity matching under viewpoint changes, illumination variation, and partial occlusion [54,55].

In MOT, most association strategies rely primarily on geometric, motion-based, and appearance-based evidence.

Some approaches rely solely on spatial overlap between bounding boxes [5,6] for association to determine matched detections with tracklet estimations. This paradigm fails under heavy occlusion conditions and some methods have been proposed to improve its performance [56,57]. In formulations such as DeepSORT [10] and StrongSORT [11] association is resolved by combining spatial overlap with motion prediction and/or learned appearance descriptors computed from detection crops.

All these design paradigms can benefit from the exploitation of semantic attributes which may provide an additional source of consistency, especially in ambiguous association cases. For example, two detections that are spatially close and visually similar but have incompatible apparent demographic predictions are less likely to correspond to the same individual than two detections that are also consistent in age group and gender. A sudden apparent age-group or gender mismatch can be treated as weak evidence against an assignment, especially when spatial and appearance cues are ambiguous. The use of semantic and attribute cues for detection-to-track association extends the classical IoU-based and appearance-based matching criteria of tracking-by-detection pipelines with higher-level semantic consistency constraints. Augmenting the association cost matrix with attribute consistency terms can therefore provide additional evidence against unlikely assignments. This observation is particularly relevant in privacy-aware retail tracking, where face information is unavailable and body-level descriptors may be affected by low resolution, occlusion, and similar clothing.

The present work follows this association-level perspective. Rather than training a joint tracking-classification model, we incorporate apparent age-group and gender predictions into the detection-to-tracklet association cost. In this way, demographic predictions are used as weak semantic consistency cues that complement IoU-based spatial matching, LSTM-based motion forecasting, and ReID appearance embeddings. This provides a lightweight and modular alternative to fully joint multi-task architectures, while directly targeting identity preservation in retail tracking scenarios.

3. Proposed Method

Figure 1 summarizes the proposed pipeline and illustrates how spatial, motion, appearance, and demographic cues are integrated before the final association stage.

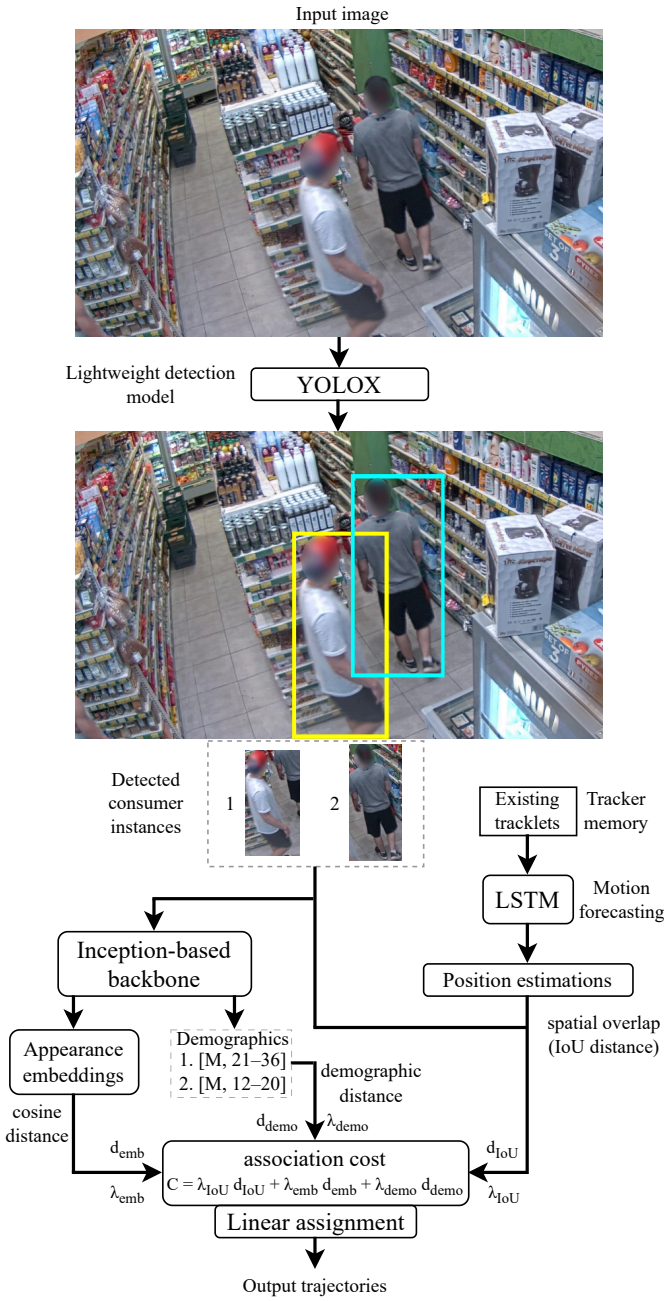

Figure 1. Overview of the proposed demographic-aware MOT pipeline. YOLOX detections, LSTM-based motion cues, ReID appearance embeddings, and apparent demographic predictions are combined within the association cost to improve consumer trajectory consistency.

3.1. Problem Formulation

The tracking module is responsible for assigning detections across frames and constructing a separate trajectory for each consumer. A trajectory can be represented as a sequence of bounding boxes:

$$T = \{B_t, \ldots, B_{t+k}\}, \quad B_i \in \mathbb{R}^4, \tag{1}$$

where each bounding box is defined as:

$$B_i = (x, y, w, h), \tag{2}$$

with x and y denoting the spatial position of the bounding box, while w and h correspond to its width and height. The goal of a tracking algorithm is the generation of a set of sequences $\hat{T}$ that represent the movement trajectories of all object instances within the tracked scene.

3.2. Multi-Object-Tracking Framework

The proposed methodology adopts the tracking-by-detection paradigm, where individuals are first detected independently in each video frame and are then associated over time in order to form continuous trajectories. In this framework, the detection stage provides bounding-box candidates for consumers appearing in the scene, while the tracking stage is responsible for maintaining their identities across consecutive frames. The overall pipeline is modular and consists of three main components: an object detection network, a motion prediction module, and an association mechanism. The detection network extracts consumer bounding boxes together with confidence scores for each frame. These detections are then passed to the tracking module, which associates them with existing trajectories. To improve the robustness of the association process, especially in cases of temporary occlusions or irregular consumer motion, motion forecasting is incorporated through an LSTM-based model, while individual consumer appearance embeddings and demographic information are extracted using a CNN-based Re-ID module.

Each module is trained separately and integrated into a unified tracking pipeline. This modular design allows the tracker to combine detection information with temporal motion cues, appearance embeddings, and demographic consistency cues, improving trajectory continuity in challenging indoor retail environments without requiring end-to-end retraining of all components. The final output of the system is a set of trajectories, where each trajectory corresponds to a detected consumer and describes their movement throughout the video sequence.

3.3. Object Detection

To identify consumer instances that appear in a scene, this work employs the YOLOX-S detection model [58]. Its architecture adopts a feature pyramid backbone with an anchor-free design that uses decoupled heads for classification and bounding box regression. This structure supports both high accuracy and computational efficiency. For every frame, the detector produces bounding boxes along with confidence scores, which are then passed to the tracking stage. To ensure a fair comparison, all evaluated methods for tracking rely on the same detection backbone.

3.4. Object Tracking

At every frame, the tracker attempts to associate the newly detected bounding boxes with already existing tracklets. These tracklets may be categorized into different states depending on their recent matching history, such as active, inactive, or lost. Active tracklets correspond to targets that have been successfully matched in recent frames, while inactive or lost tracklets refer to targets that have temporarily disappeared, usually due to occlusion, missed detections, or exiting the visible region of the scene.

The association procedure follows a detector-centered strategy. Initially, detections are divided according to their confidence scores. High-confidence detections are matched first with the predicted locations of existing tracklets. For this purpose, a cost matrix is constructed using the intersection-over-union metric between detections and predicted tracklet positions. The optimal assignment is then obtained through the Hungarian algorithm.

After the first association step, unmatched tracklets and detections are processed in additional matching stages. Lower-confidence detections may still contain valid consumer observations and are therefore considered in a subsequent association step. This allows the tracker to recover targets that would otherwise be ignored due to lower detector confidence. In addition, inactive tracklets are also considered in the matching process, enabling the continuation of trajectories after short-term occlusions or detection failures.

When a detection cannot be matched with any existing tracklet, a new trajectory is initialized. Conversely, if a tracklet remains unmatched for a predefined number of frames, it is first marked as lost and may eventually be removed from the tracking process. This strategy allows the tracker to maintain stable identities while avoiding the indefinite preservation of trajectories that are no longer visible in the scene.

3.5. Demographic-Aware Association Cost

Let T_i denote an active tracklet and D_j^t a detection at frame t. The association cost between active tracklets and detections obtained from the detection network is defined as

$$C_{ij}^t = d_{\text{IoU}}(T_i, D_j^t) \tag{3}$$

where d_{IoU} is measured as the spatial IoU overlap between two bounding boxes.

We propose an association cost that incorporates appearance feature embeddings and apparent demographic predictions extracted from the Inception-based demographic estimation module:

$$C_{ij}^t = d_{\text{IoU}}(T_i, D_j^t) + d_{\text{emb}}(T_i, D_j^t) + d_{\text{demo}}(T_i, D_j^t), \tag{4}$$

where

- d_{emb} denotes the appearance embedding distance,
- d_{demo} represents the demographic distance.

The appearance embedding distance follows the cosine rule:

$$1 - \frac{e_{T_i} \cdot e_{D_j^t}}{||e_{T_i}||_2 ||e_{D_j^t}||_2}, \tag{5}$$

while the demographic distance is defined as

$$d_{\text{demo}} = w_g \cdot \mathbb{1}(g_i \neq g_j) + w_a \cdot \frac{|a_i - a_j|}{K}, \tag{6}$$

where

- g_i, g_j are gender labels,
- a_i, a_j are ordinal age-group indices,
- K is the maximum ordinal distance, and
- w_g, w_a are weighing factors to balance the impact of each attribute, set at 0.25 and 0.75, respectively.

To avoid instances where one factor dominates the association cost matrix, which can lead to missed associations, we weigh each distance by a factor λ_i:

$$C_{ij}^t = \lambda_{\text{IoU}} d_{\text{IoU}}(T_i, D_j^t) + \lambda_{\text{emb}} d_{\text{emb}}(T_i, D_j^t) + \lambda_{\text{demo}} d_{\text{demo}}(T_i, D_j^t), \tag{7}$$

In practice, we find that d_{IoU} is the most reliable distance metric, due to the difficulty of accurately estimating the gender and age group of individuals, followed by d_{ReID}. For this reason, we set λ_{IoU} to 0.75, λ_{ReID} to 0.2 and λ_{demo} to 0.05. In addition, since the overall

weighing factor for d_{demo} is set to 0.05, different values of w_g and w_a have no effect on overall performance in our experiments. An ablative analysis of the weighing factors for each association cost term is provided in Section 5.2.

4. Experimental Setup

4.1. Dataset

In this work, we employ the *Consumers* dataset [3], which comprises sequences of consumers in indoor retail scenes. The sequences were recorded under normal store operation and depict individuals in realistic scenarios, where variations in external appearance (e.g., clothing, accessories), occlusions between consumers and the environment and irregular movement patterns are common. Additionally, each sequence contains one or more consumers entering or leaving the scene and interactions between individuals are frequent. In contrast to standard pedestrian tracking benchmarks, consumer motion in retail environments is often less regular, since individuals may stop near shelves, change direction suddenly, or interact with other consumers and objects. These characteristics make the dataset particularly suitable for evaluating tracking methods in realistic retail scenarios.

The dataset comprises a total of 8700 images of spatial resolution 3072×1729 in 145 sequences of varying length, organized as groups of images, of which 120 are used as the training set, and the remaining 25 as the test set. The images belonging to the training set are used to train the detection models, while the images of the test set are used to validate both the detection models as well as the trackers. In the test set, there are a total number of 1134 objects to be detected and tracked across 49 unique trajectories, split among the 25 sequences. An average of 58.64 frames corresponds to each sequence, with 1466 in total, which amounts to 0.77 objects per frame since there are instances of scenes with no visible consumers such as before and after exiting the camera field of view or during occlusion due to the environment. The original dataset splits [3] are followed to ensure a fair comparison between the different tracking methods. All compared tracking methods are evaluated under the same detection configuration. In this way, the comparison focuses on the effectiveness of the tracking and association strategies rather than differences caused by the detector.

An additional challenge of this dataset is the presence of anonymization in the form of blurring of the individuals' face area, to adhere to privacy-respecting regulations such as GDPR. This introduces compounding difficulties for the detection and Re-ID models, especially for the latter, since a section of the image that is commonly associated with gender (head area and hair) is blurred, hampering the accuracy of estimations. The Re-ID model is trained on the *BID* [3] dataset, where the train and test set images follow the same split as in *Consumers*, since they are the cropped detections. This way, there is a clear distinction between train and test sets, for both detection and Re-ID models. A statistical breakdown of the *BID* dataset is included in Table 1.

Table 1. A statistical breakdown of the *BID* dataset, for the train and test splits of each demographic attribute.

Split	**Samples**	**Gender Attribute**			
		M		**F**	
train	5507	39.2%		60.8%	
test	1134	44.9%		55.1%	
Split	**Samples**	**Age Group Attribute**			
		12–20	21–36	37–60	61+
train	5507	5.4%	50.4%	39.6%	4.6%
test	1134	12.8%	42.5%	38.7%	6.0%

4.2. Training and Implementation Details

All models are implemented using the PyTorch 2.4.1 framework and are trained separately using a single GPU. After training they are unified in the tracking framework.

The detection model is trained on the *Consumers* training set in the task of person detection. We use the provided weights of [3] as baseline and fine-tune the network for an additional 10 epochs with a 0.001 learning rate using Stochastic Gradient Descent. The first epoch is used as a warm-up period, and the learning rate is annealed using a cosine scheduler. All other training settings used follow the defaults set by [58]. We train several lightweight models intended for mobile applications, using two versions of the *Consumers* dataset. In the private, held-out version, the images used are raw without any form of anonymization applied; while in the second, public version, the area around the face of each consumer has been blurred in order to respect the privacy of each individual.

The LSTM network follows the training procedure of [59], where the training split (7 sequences) of the *MOT17* dataset is employed for a total of 100 epochs. For this model, the Adam optimizer with an initial learning rate of 0.001 is used, and after 60 epochs we reduce the learning rate to 0.0001. In our implementation, the LSTM uses a single cell with 128 neurons, as larger configurations did not provide a significant benefit in overall tracking quality. Since the LSTM is utilized at each tracker update step for every tracklet, applications in scenarios with a large number of targets can potentially cause an increase in network latency due to the added computations. By using a lightweight LSTM module, the overhead of the framework is kept at manageable levels, while the tracking algorithm benefits from its inclusion.

Finally, the Re-ID model adopts the Inception-based architecture of [60], is initialized with *ImageNet* weights and fine-tuned on the *BID* [3] dataset. *BID* comprises full-body images of consumers where the area near the subject's face has been blurred to preserve the privacy of individuals and to comply with data regulations. Due to the relatively small sample size of the dataset, we use a learning rate of 0.0001 with a weight decay of 0.0005 for 10 epochs, to avoid overfitting. That architecture has demonstrated acceptable performance for indoor full-body image demographic estimation [3] and is modified to also extract a 1024-dimensional feature vector from its input, which is used by the tracker during the stage of tracklet association.

Conventional frame-level k-fold cross-validation was not employed because the video data consist of temporally correlated sequences, and a random partition of individual frames could introduce temporal leakage between the training and validation subsets. We therefore retained the original sequence-level training and evaluation split adopted in [3], ensuring direct comparability with the previously reported results. A separate sequence-level cross-validation protocol was not introduced in the present study. Since each trainable model was optimized once under this fixed split, the reported results do not include repeated-seed uncertainty estimates; this constitutes a limitation of the present evaluation.

5. Results and Discussion

5.1. Quantitative Results

For evaluation of our proposed method, we adopt the privacy-aware protocol presented in [15], which involves two evaluation experiments simulating realistic conditions where information in consumers' faces has been blurred to preserve their identity following privacy-respecting regulations. The first experiment trains the detection network using the raw *Consumers* training set, while in the second we use the publicly released *Consumers* training set which contains sequences where blurring has been applied to the areas of the

face. This scenario also represents a realistic use case where raw data is not always available due to external constraints.

Our proposed method is compared against several approaches from the multi-object tracking literature and we report the following evaluation indices to assess tracking quality:

- the *CLEAR MOT* [61] metrics: MOTA and MOTP,
- the tracklet identity metric IDF1,
- the tracking quality indicators: mostly-tracked (MT), partially-tracked (PT) and mostly-lost (MT), and
- the amount of identity switches that occurred.

For a fair comparison, all methods are evaluated using the same weights obtained from the training procedure outlined in the previous Section using identical training hyperparameters. The tracking results of each experiment are presented in Tables 2 and 3.

Table 2. Quantitative results of several MOT methods on the *Consumers* test set with the object detection method trained on the original raw data. All methods use the same weights for the detection model. Reported results are the average of the 25 sequences of the test set. Best results are highlighted in bold. ↑ indicates that a higher score is better, ↓ the opposite.

Method	Association Type *	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	PT	ML ↓	IDSw ↓
SORT [6]	S	75.0%	81.3%	83.2%	27	20	**2**	13
ByteTrack [5]	S	65.2%	81.3%	62.2%	26	21	**2**	115
DeepSORT [10]	S + A	54.1%	76.4%	69.4%	16	31	**2**	22
StrongSORT [11]	S + A	75.0%	86.8%	80.9%	28	19	**2**	15
OC-SORT [56]	S	75.1%	**88.5%**	81.8%	28	19	**2**	8
PD-SORT [57]	S	78.4%	87.8%	80.3%	24	20	5	8
Byte + LSTM [3]	S	83.6%	88.4%	87.7%	37	10	**2**	**6**
Byte + LSTM + ReID [15]	S + A	84.4%	87.5%	87.2%	40	7	**2**	**6**
Ours	S + A + D	**84.8%**	87.8%	**87.8%**	**42**	5	**2**	**6**

* S: spatial; A: appearance; D: demographic.

The experimental assessment shows that incorporating demographic attributes, namely apparent gender and age-group predictions, into the tracklet association process provides consistent benefits over the compared MOT configurations. In the raw-data setting, the proposed method achieves the highest MOTA and IDF1 scores among all evaluated trackers, while also increasing the number of mostly tracked trajectories. Compared with the strongest internal baseline, ByteTrack + LSTM + Re-ID [15], the proposed demographic-aware association improves MOTA, MOTP, IDF1, and mostly tracked trajectories, while preserving the same low number of identity switches. This indicates that demographic consistency does not replace spatial, motion, or appearance cues, but acts as an additional semantic constraint that improves the final association decision.

Simple spatial overlap association strategies such as SORT and ByteTrack are outperformed in the *Consumers* dataset by newer approaches such as OC-SORT and PD-SORT due to instances of consumer occlusion and irregular movement, which Kalman-based motion estimation fails to accurately predict. This is further supported by [3], where replacing the Kalman filter with an LSTM network provides a significant boost in performance, from 63.1% MOTA to 77.1%, higher than the aforementioned methods. ReID-based association trackers such as DeepSORT and StrongSORT obtain widely different tracking scores, which can be attributed to the simpler architecture and ReID model employed by DeepSORT that performs sub-optimally in this dataset. Due to being similar algorithms, OC-SORT and PD-SORT both demonstrate strong MOTP scores, with the former actually

being the highest in our comparison, suggesting a potential application of these methods in retail scenarios.

Table 3. Quantitative results of several MOT methods on the *Consumers* test set with the object detection method trained on the public version of the dataset, whereby individual information of consumers has been blurred. All methods use the same weights for the detection model. Reported results are the average of the 25 sequences of the test set. Best results are highlighted in bold. ↑ indicates that a higher score is better, ↓ the opposite.

Method	Association Type *	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	PT	ML ↓	IDSw ↓
SORT [6]	S	68.4%	80.1%	77.5%	20	25	4	11
ByteTrack [5]	S	63.1%	81.6%	62.4%	25	20	4	95
DeepSORT [10]	S + A	58.2%	75.5%	72.1%	17	28	4	14
StrongSORT [11]	S + A	71.1%	85.4%	81.2%	25	19	5	7
OC-SORT [56]	S	73.6%	**87.8%**	80.3%	24	20	5	8
PD-SORT [57]	S	74.5%	87.0%	79.4%	22	20	7	8
Byte + LSTM [3]	S	77.1%	85.6%	81.0%	30	15	4	6
Byte + LSTM + ReID [15]	S + A	79.0%	86.3%	84.0%	34	**10**	5	**2**
Ours	S + A + D	**80.0%**	87.6%	**85.3%**	**34**	12	**3**	**2**

* S: spatial; A: appearance; D: demographic.

A similar trend is observed under the privacy-aware experimental protocol, where the detector is trained on the public anonymized version of the *Consumers* dataset. In this setting, the proposed method again achieves the best MOTA, MOTP, and IDF1 scores, while maintaining the same number of identity switches as the strongest Re-ID-based variant and reducing the number of mostly lost trajectories. This result is particularly relevant because face blurring introduces visual degradation in the head region, which can affect both detection and appearance-based association. Since all compared methods use the same detection model and the same detector weights within each experiment, the observed differences are attributed to the association strategy rather than to differences in detection quality.

In a nutshell, the results suggest that even a conservative demographic-aware term can improve the tracking process when used as a weak semantic cue within the association cost. The proposed method does not require a new demographic classifier or an end-to-end tracking-classification architecture. On the contrary, it reuses existing apparent demographic predictions to provide additional identity-consistency evidence, leading to incremental but stable improvements in tracking performance under both raw and anonymized evaluation conditions.

5.2. Hyperparameter Analysis

Our proposed association cost employs three separate cues during calculation: spatial bounding box overlap, appearance embeddings and demographic estimations for each detected consumer instance. Each particular cue has a weak point, e.g., bounding box overlap is hampered by occlusions, and relying exclusively on a single cue during association could be detrimental to final performance, especially in cases where that weakness is present. For this reason, we apply a weighing factor to each cue, balancing the contribution of each term to the overall cost calculation, which ultimately affects the final identity assignment. In Table 4, we showcase results of our proposed method for different values of the three weight hyperparameters used to balance the cost association terms.

Table 4. Hyperparameter analysis of the association cost term weighing factors. In this experiment, the same object detection model is used and evaluation is performed on the anonymized *Consumers* test set. Reported results are the average of the 25 sequences of the test set. Best results are highlighted in bold. ↑ indicates that a higher score is better, ↓ the opposite.

λ_{IoU}	λ_{emb}	λ_{demo}	MOTA ↑	MOTP ↑	IDF1 ↑	MT ↑	PT	ML ↓	IDSw ↓
0.75	0.2	0.05	**80.0%**	**87.6%**	**85.3%**	**34**	12	**3**	**2**
0.75	0.05	0.2	79.8%	87.4%	85.2%	33	11	5	**2**
0.5	0.25	0.25	79.5%	87.2%	85.0%	33	10	6	**2**
0.8	0.1	0.1	**80.0%**	**87.6%**	**85.3%**	**34**	12	3	**2**
0.9	0.1	0.1	79.9%	**87.6%**	85.2%	**34**	11	4	**2**

In practice, we find that the proposed method is robust to a varying selection of hyperparameters, with bounding box overlap being the most significant factor for the *Consumers* dataset. In fact, exploiting the memory of LSTM allows the framework to maintain high tracking scores without overly relying on a particular association term; however, the best tracking performance is achieved for a hyperparameter selection that emphasizes bounding box overlap rather than appearance embeddings or demographic attributes. Moreover, our results indicate that favoring appearance embeddings over demographic information with a higher weight can benefit the final results, especially in cases where the demographic estimation is unreliable due to high uncertainty and changes in consecutive frames, e.g., due to consumer occlusion or rotation. Notably, while we can set the total value of the weighing terms (i.e., $\sum \lambda$) to <1 or >1, we find that doing so negatively affects tracking quality by leading to association costs that penalize tracklet-to-detection matching, and for this reason we choose values that sum to a maximum of 1.

5.3. Runtime Analysis

To support deployment in practical environments with realistic scenarios where strict limits are often imposed on computational resources, we additionally analyze the efficiency of our approach, reporting the associated overhead in overall latency. Our runtime evaluation analyzes the end-to-end execution speed of our proposed method using the lightweight model variants of the YOLOX series [58] to assess the scalability of our end-to-end approach under constrained runtime budgets. Furthermore, in order to simulate real-world conditions where the privacy of individuals is an additional limitation, the detection models are trained and evaluated on the public *Consumers* dataset using privacy-compliant data, i.e., blurring over the area of the face.

The runtime measurements are obtained on a machine running Linux 18.04.5 LTS using a NVIDIA RTX 2080 Ti GPU and an Intel Xeon processor running in single core mode. For each detection model, 10 evaluation runs in the *Consumers* testing sequences are performed and the average FPS are calculated. We report MOTA, MOTP and IDF1 scores along with the mean and standard deviation of the latency measurements in Table 5.

Our runtime analysis showcases a trade-off between latency and performance, especially in the case of YOLOX-S vs. Tiny models. While YOLOX-S outperforms the other detection models with more accurate detections as reflected in the tracking metrics, its latency falls behind due to its larger model size and computational overhead. The latency of YOLOX-Tiny lies between the other two models and its tracking performance follows closely that of YOLOX-S, demonstrating a compromise between tracking quality and overall latency. Finally, YOLOX-Nano achieves the highest throughput at 22.74 FPS, compared to 18.25 FPS for YOLOX-S, but at the cost of lower tracking performance. This corresponds to an increase of 4.49 FPS, or approximately 24.6%, over YOLOX-S in the evaluated setup. The measured runtime behavior may also be influenced by the use of

depthwise convolutions, whose practical efficiency depends on the underlying hardware and software implementation.

Table 5. Runtime and performance analysis of our proposed method with different lightweight detection methods with the YOLOX architecture. Evaluations are performed in the public anonymized *Consumers* test set. In each experiment, the Re-ID model used follows the Inception-based architecture described in Section 3. Model size and FLOPS refer to the combined architecture (detection and Re-ID). Parameter size is measured in millions. Best results are highlighted in bold. ↑ indicates that a higher score is better.

Detection Model	Params	GFLOPS	MOTA ↑	MOTP ↑	IDF1 ↑	FPS ↑
YOLOX-S	20.39	28.2	**80.0%**	**87.6%**	**85.3%**	18.25 ± 1.0
YOLOX-Tiny	16.99	7.8	79.5%	86.9%	85.2%	22.03 ± 0.9
YOLOX-Nano	12.30	2.5	78.2%	84.7%	83.9%	**22.74 ± 0.5**

5.4. Limitations and Future Work

The present study also defines several directions for further improvement. The demographic-aware term is deliberately introduced with a conservative weight, so that unreliable age-group or gender predictions cannot dominate the association decision. This design choice preserves the robustness of the tracker, but also suggests that more adaptive weighting strategies could further improve performance. In addition, the current weighting factors are manually selected to keep the association model simple and interpretable. Learning these weights from validation data, or adapting them according to the confidence of the detector, ReID module, and demographic classifier, could provide a more flexible formulation. Finally, the evaluation focuses on RGB retail sequences from the *Consumers* dataset; broader validation across additional store layouts, camera viewpoints, crowd densities, and multi-camera settings would further establish the generality of the approach.

A promising future direction is to move from fixed hand-designed association costs toward adaptive demographic-aware association models. One possible formulation is an expectation–maximization-inspired scheme in which tracking and demographic estimation are treated as mutually informative but separately optimized processes. In one direction, stable demographic predictions could provide latent semantic consistency cues for adapting the association cost when spatial and appearance information is ambiguous. In the reverse direction, stable tracking identities could provide temporally coherent samples for improving demographic predictions over tracklets, rather than treating each frame independently. Such a formulation could be further extended with learnable demographic embeddings, confidence-adaptive weighting, or multi-expert fusion mechanisms that combine spatial, motion, appearance, and semantic cues. These extensions are beyond the scope of the present work, but they follow naturally from the modular formulation proposed here and could transform the current conservative demographic-aware cost into a more data-driven association model.

6. Conclusions

This work presented a demographic-aware multi-object tracking framework for retail environments. The proposed method follows a modular tracking-by-detection pipeline, where YOLOX-based detections are associated over time using spatial overlap, LSTM-based motion forecasting, ReID appearance embeddings, and apparent demographic cues. The main contribution of the work is the integration of age-group and gender predictions into the detection-to-tracklet association cost, allowing demographic consistency to act as a weak semantic cue for identity preservation.

The experimental evaluation on a realistic dataset demonstrated that incorporating demographic consistency can improve tracking performance in indoor retail scenes, un-

der both raw and anonymized training conditions, improving MOTA and IDF1 scores as well as tracking quality in the form of mostly tracked trajectories. These results indicate that demographic-aware association can contribute to more stable consumer trajectories, especially in scenarios where occlusions, visual similarity, and low-resolution body crops make identity preservation difficult. The privacy-aware evaluation further showed that the method remains effective when face regions are blurred, supporting its applicability to retail settings where data protection constraints limit the use of identifiable facial information.

The runtime analysis revealed the trade-off between tracking accuracy and computational efficiency when using different lightweight YOLOX detector variants, suggesting that the proposed association strategy can be integrated into practical retail analytics pipelines, although its overall efficiency remains dependent on the selected detector and the computational budget of the deployment environment.

Author Contributions: Conceptualization, I.-I.P. and A.P.G.; methodology, I.-I.P. and A.P.G.; software, I.-I.P.; validation, I.-I.P.; formal analysis, I.-I.P. and A.P.G.; investigation, I.-I.P. and A.P.G.; resources, I.-I.P. and G.G.; data curation, I.-I.P. and A.P.G.; writing—original draft preparation, I.-I.P. and A.P.G. and M.E.P. and V.S. and G.G.; writing—review and editing, I.-I.P. and A.P.G. and M.V. and C.N.; visualization, I.-I.P.; supervision, A.P.G. and M.V. and C.N.; project administration, C.N.; funding acquisition, A.P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been co-financed by the European Union–NextGenerationEU through the National Recovery and Resilience Plan "Greece 2.0", under the Action "Clusters of Research Excellence (CREs)" (Project: DEVIATE; Project Code: YP3TA-0560419; MIS: 5180519), implemented by the Executive Structure of the NSRF of the Ministry of Education, Religious Affairs and Sports.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was waived because the present study does not involve the collection of new human-subject data and relies only on previously published anonymized retail image sequences. Face regions are blurred in the public dataset, and the analysis is performed on anonymized body crops and tracking outputs without using directly identifying personal information.

Data Availability Statement: The *Consumers* and Body-Image Dataset (*BID*) used in this study is available at https://www.kaggle.com/datasets/angelosgiotis/consumers-bid (accessed on 30 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Quintana, M.; Menendez, J.; Alvarez, F.; Lopez, J. Improving retail efficiency through sensing technologies: A survey. *Pattern Recognit. Lett.* **2016**, *81*, 3–10. [CrossRef]
2. Paolanti, M.; Pietrini, R.; Mancini, A.; Frontoni, E.; Zingaretti, P. Deep understanding of shopper behaviours and interactions using RGB-D vision. *Mach. Vis. Appl.* **2020**, *31*, 66. [CrossRef]
3. Panagos, I.I.; Giotis, A.P.; Sofianopoulos, S.; Nikou, C. A New Benchmark for Consumer Visual Tracking and Apparent Demographic Estimation from RGB and Thermal Images. *Sensors* **2023**, *23*, 9510. [CrossRef] [PubMed]
4. Bergmann, P.; Meinhardt, T.; Leal-Taixé, L. Tracking Without Bells and Whistles. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 941–951.
5. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv, Israel, 23–27 October 2022; pp. 1–21.
6. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.T.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
7. Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; Xia, H. End-to-End Video Instance Segmentation with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 19–25 June 2021; pp. 8737–8746. [CrossRef]

8. Meinhardt, T.; Kirillov, A.; Leal-Taixé, L.; Feichtenhofer, C. TrackFormer: Multi-Object Tracking with Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 8834–8844. [CrossRef]
9. Hou, M.; Wu, Y.; Shi, H.; Mu, X. A Two Stage Multi Object Tracking Algorithm with Transformer and Attention Mechanism. *Sci. Rep.* **2025**, *15*, 31414. [CrossRef] [PubMed]
10. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. [CrossRef]
11. Du, Y.; Zhao, Z.; Song, Y.; Zhao, Y.; Su, F.; Gong, T.; Meng, H. StrongSORT: Make DeepSORT Great Again. *IEEE Trans. Multimed.* **2023**, *25*, 8725–8737. [CrossRef]
12. Li, Y.; Zhan, L.; Wu, L.; Liu, H.; Min, J.; Wang, X. Re-identification Assistance and Multi-stage Association for Pedestrian Multi-object Tracking. *Sci. Rep.* **2025**, *15*, 22533. [CrossRef] [PubMed]
13. Xu, J.; Cao, Y.; Zhang, Z.; Hu, H. Spatial-Temporal Relation Networks for Multi-Object Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3987–3997. [CrossRef]
14. Ma, F.; Shou, M.Z.; Zhu, L.; Fan, H.; Xu, Y.; Yang, Y.; Yan, Z. Unified Transformer Tracker for Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 8781–8790. [CrossRef]
15. Panagos, I.I.; Giotis, A.P.; Stergiou, V.; Voudiotis, G.; Manos, A.; Filippidou, D.E.; Nikou, C. Privacy-Aware Multi-Object Tracking for Retail Environments. In Proceedings of the 18th International Conference on Digital Image Processing (ICDIP 2026), Xiamen, China, 24–26 April 2026.
16. Leykin, A.; Tuceryan, M. *Tracking and Activity Analysis in Retail Environments*; Indiana University Computer Science Technical Reports; Indiana University: Bloomington, IN, USA, 2005.
17. Cobos, R.; Hernandez, J.; Abad, A.G. Retail Traffic-Flow Analysis Using a Fast Multi-object Detection and Tracking System. *Appl. Comput. Intell.* **2019**, *1096*, 29–39. [CrossRef]
18. Jose, J.A.C.; Bertumen, C.J.B.; Roque, M.T.C.; Umali, A.E.B.; Villanueva, J.C.T.; TanAi, R.J.; Sybingco, E.; San Juan, J.; Gonzales, E.C. Smart Shelf System for Customer Behavior Tracking in Supermarkets. *Sensors* **2024**, *24*, 367. [CrossRef] [PubMed]
19. Nguyen, T.D.; Hihara, K.; Hoang, T.C.; Utada, Y.; Torii, A.; Izumi, N.; Thuy, N.T.; Pham, D.T.; Tran, L.Q. Retail Store Customer Behavior Analysis System: Design and Implementation. In Proceedings of the Artificial Intelligence Applications and Innovations, Corfu, Greece, 27–30 June 2024; pp. 305–318. [CrossRef]
20. Shili, M.; Jayasingh, S.; Hammedi, S. Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment. *Electronics* **2024**, *13*, 4730. [CrossRef]
21. Mendes, D.; Correia, S.; Jorge, P.; Brandão, T.; Arriaga, P.; Nunes, L. Multi-Camera Person Re-Identification Based on Trajectory Data. *Appl. Sci.* **2023**, *13*, 11578. [CrossRef]
22. Nakahara, T.; Yada, K. Analyzing consumers' shopping behavior using RFID data and pattern mining. *Adv. Data Anal. Classif.* **2012**, *6*, 355–365. [CrossRef]
23. Fujino, T.; Kitazawa, M.; Yamada, T.; Takahashi, M.; Yamamoto, G.; Yoshikawa, A.; Terano, T. Analyzing In-Store Shopping Paths from Indirect Observation with RFID Tags Communication Data. *RISUS J. Innov. Sustain.* **2014**, *5*, 88–96. [CrossRef]
24. Ali, K.; Liu, A.X.; Chai, E.; Sundaresan, K. Monitoring Browsing Behavior of Customers in Retail Stores via RFID Imaging. *IEEE Trans. Mob. Comput.* **2022**, *21*, 1034–1048. [CrossRef]
25. Sturari, M.; Liciotti, D.; Pierdicca, R.; Frontoni, E.; Mancini, A.; Contigiani, M.; Zingaretti, P. Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. *Pattern Recognit. Lett.* **2016**, *81*, 30–40. [CrossRef]
26. Stavrou, V.; Papakyriakopoulos, D. Smart Retail: Efficient in-store localization using ensemble classifiers. In Proceedings of the 24th Pan-Hellenic Conference on Informatics (PCI '20), Athens, Greece, 20–22 November 2020; pp. 301–304. [CrossRef]
27. Ayaz, Z. Digital Advertising and Customer Movement Analysis Using BLE Beacon Technology and Smart Shopping Carts in Retail. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 55. [CrossRef]
28. Panagos, I.I.; Giotis, A.P.; Nikou, C. Tracking Multiple Instances of Retail Consumers from RGB and Thermal Images. *Eng. Proc.* **2022**, *21*, 17. [CrossRef]
29. Ahmar, W.E.; Massoud, Y.; Kolhatkar, D.; Alghamdi, H.; Alja'Afreh, M.; Laganiere, R.; Hammoud, R. Enhanced Thermal-RGB Fusion for Robust Object Detection. In Proceedings of the IEEE/CVF Conference Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2023. [CrossRef]
30. Qin, Y.; Zhang, J.; Fan, S.; Liu, Z.; Wang, J. MCIT: Multi-level cross-modal interactive transformer for RGBT tracking. *Neurocomputing* **2025**, *649*, 130758. [CrossRef]
31. Wang, D.; Zhuang, L.; Gao, L.; Sun, X.; Zhao, X. Global feature-injected blind-spot network for hyperspectral anomaly detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5509305. [CrossRef]

32. Teng, X.; Zhao, D.; Xiang, P.; Yu, X.; Li, Z.; Hsu, C.C.; Yang, T.; Zhou, H.; Ren, J. UAV-based multispectral object tracking with positive–negative prompt mining network. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 5515418. [CrossRef]
33. Çiftçi, S.; Akyüz, A.O.; Ebrahimi, T. A Reliable and Reversible Image Privacy Protection Based on False Colors. *IEEE Trans. Multimed.* **2018**, *20*, 68–81. [CrossRef]
34. Wu, H.; Tian, X.; Li, M.; Liu, Y.; Ananthanarayanan, G.; Xu, F.; Zhong, S. PECAM: Privacy-enhanced video streaming and analytics via securely-reversible transformation. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom'21), New Orleans, LA, USA, 25–29 October 2021; pp. 229–241. [CrossRef]
35. Sun, Q.; Ma, L.; Joon Oh, S.; Gool, L.V.; Schiele, B.; Fritz, M. Natural and Effective Obfuscation by Head Inpainting. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5050–5059. [CrossRef]
36. Megala, G.; Venkatesan, R.; Vigneshwaran, T. Video Face Anonymization for Preserving Privacy. In Proceedings of the 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), Chennai, India, 8–10 December 2022; Volume 1, pp. 1–5. [CrossRef]
37. Rosberg, F.; Aksoy, E.; Englund, C.; Alonso-Fernandez, F. FIVA : Facial Image and Video Anonymization and Anonymization Defense. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023; pp. 362–371. [CrossRef]
38. Simões, V.; Garcia, J.; Carreira, P. Privacy-aware video deanonymization: A configurable pipeline for selective reversal with intelligibility preservation. *J. Inf. Secur. Appl.* **2026**, *97*, 104367. [CrossRef]
39. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards Realtime Multi-Object Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual, 23–28 August 2020. [CrossRef]
40. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
41. Chan, S.; Qiu, C.; Wu, D.; Hu, J.; Heidari, A.A.; Chen, H. Fusion detection and ReID embedding with hybrid attention for multi-object tracking. *Neurocomputing* **2024**, *575*, 127328. [CrossRef]
42. Wang, H.; Chen, T.; Wang, Y. Towards Occlusion-Aware Multi-Pedestrian Tracking. *Appl. Sci.* **2025**, *15*, 13045. [CrossRef]
43. Bayraktar, E. ReTrackVLM: Transformer-Enhanced Multi-Object Tracking with Cross-Modal Embeddings and Zero-Shot Re-Identification Integration. *Appl. Sci.* **2025**, *15*, 1907. [CrossRef]
44. Yuan, C.; Guan, Z.; Xu, Y.; Zhu, X.; Liang, W. Jointly Adaptive Cross-Resolution Person Re-Identification on Super-Resolution. *Complex Intell. Syst.* **2025**, *11*, 249. [CrossRef]
45. Ranjan, R.; Zhou, S.; Cheng Chen, J.; Kumar, A.; Alavi, A.; Patel, V.M.; Chellappa, R. Unconstrained Age Estimation With Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Santiago, Chile, 7–13 December 2015; pp. 351–359. [CrossRef]
46. Al-Shannaq, A.S.; Elrefaei, L.A. Comprehensive Analysis of the Literature for Age Estimation From Facial Images. *IEEE Access* **2019**, *7*, 93229–93249. [CrossRef]
47. Sharma, N.; Sharma, R.; Jindal, N. Face-Based Age and Gender Estimation Using Improved Convolutional Neural Network Approach. *Wirel. Pers. Commun.* **2022**, *124*, 3035–3054. [CrossRef]
48. Ge, Y.; Lu, J.; Fan, W.; Yang, D. Age estimation from human body images. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 2337–2341. [CrossRef]
49. Dai, B.; Yang, D. Gender Classification Based on Body Images. In Proceedings of the 2021 China Automation Congress (CAC), Beijing, China, 22–24 October 2021; pp. 5549–5552. [CrossRef]
50. Xu, C.; Makihara, Y.; Liao, R.; Niitsuma, H.; Li, X.; Yagi, Y.; Lu, J. Real-Time Gait-Based Age Estimation and Gender Classification from a Single Image. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021; pp. 3459–3469. [CrossRef]
51. Yuan, B.; Wu, A.; Zheng, W.S. Does A Body Image Tell Age? In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2142–2147. [CrossRef]
52. Layne, R.; Hospedales, T.M.; Gong, S. Attributes-Based Re-identification. In *Proceedings of the Person Re-Identification*; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
53. Moussi, F.; Ouamane, A.; Ouafi, A. Person re-identification using soft biometrics. *Signal Image Video Process.* **2024**, *18*, 5599–5607. [CrossRef]
54. Li, D.; Chen, X.; Huang, K. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 111–115. [CrossRef]
55. Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; Wang, L. Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 17689–17699. [CrossRef]

56. Cao, J.; Weng, X.; Khirodkar, R.; Pang, J.; Kitani, K. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 9686–9696. [CrossRef]
57. Wang, Y.; Zhang, D.; Li, R.; Zheng, Z.; Li, M. PD-SORT: Occlusion-Robust multi-object tracking using pseudo-depth cues. *IEEE Trans. Consum. Electron.* **2025**, *71*, 3541839. [CrossRef]
58. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
59. Chaabane, M.; Zhang, P.; Beveridge, J.R.; O'Hara, S. Deft: Detection embeddings for tracking. In Proceedings of the Computer Vision and Pattern Recognition ADP3 Workshop on Autonomous Driving: Perception, Prediction and Planning, Nashville, TN, USA, 19–25 June 2021.
60. Tang, C.; Sheng, L.; Zhang, Z.X.; Hu, X. Improving Pedestrian Attribute Recognition With Weakly-Supervised Multi-Scale Attribute-Specific Localization. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4996–5005. [CrossRef]
61. Bernardin, K.; Stiefelhagen, R. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]