Abandoned Bag Detection in Public Areas Using Grounding DINO With Fine-Grained Prompts

Tomorn Soontornnapar
Department of Electrical Engineering
Siam University
Bangkok, Thailand
tomorn.soo@siam.edu

***Abstract*—Detecting abandoned bags in public areas is essential for ensuring safety and preventing potential threats. This research presents a novel approach for abandoned bag detection using Grounding DINO, a state-of-the-art vision-language model, combined with fine-grained semantic prompts. The method integrates object detection and contextual analysis to identify stationary bags left unattended for specified durations, distinguishing them from routine activities in dynamic settings such as airports, train stations, and shopping malls. The proposed approach is evaluated on the ABODA dataset, achieving an abandoned bag detection recall of 90.91%. Grounding DINO's capability to process textual prompts ensures precise bag identification, while its adaptability supports diverse public environments. The workflow involves detecting bags using fine-grained prompts, tracking their movement across frames through centroid distances between the owner's and bag's bounding boxes, and applying temporal and frame-based criteria to confirm abandonment. To enhance reliability, the system incorporates proximity-based owner identification by detecting nearby individuals and analyzing their interactions. Context-aware thresholds adjust detection parameters, ensuring robustness in crowded or complex environments. Furthermore, the results are compared to previous works to evaluate differences in performance and capabilities.**

Keywords—abandoned object detection, unattended bag, Grounding DINO, prompts, foundation model

I. Introduction

The detection of abandoned objects (AOD) in public spaces is a pivotal aspect of ensuring public safety and preventing potential threats such as terrorism, theft, and accidents. As public spaces become increasingly crowded, the reliance on automated video surveillance systems has grown, with the primary goal of minimizing human intervention while maintaining high accuracy in detecting suspicious activities. In Figure 1, the sequence illustrates a typical scenario for abandoned object detection (AOD). The detection process involves identifying specific behaviors in the video feed, such as a person carrying an object (①), placing it at a certain distance (②), walking away from the object (③), and leaving the object stationary and unattended (④). Automated systems rely on advanced algorithms to analyze such patterns by tracking the motion of individuals and their belongings, assessing temporal and spatial relationships between objects and people. This enables the system to distinguish between scenarios where an item is intentionally abandoned and situations such as temporary placement. For instance, Fig. 1 demonstrates the importance of trajectory analysis in recognizing that the person has moved away from the bag. Combined with object detection models, these systems

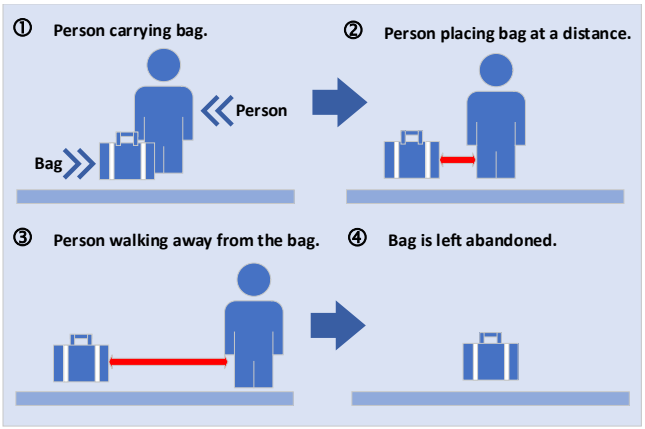

Fig. 1. Sequence depicting a person carrying a bag, placing it at a distance, walking away, and leaving the bag abandoned.

classify the bag as "abandoned" when it remains static without any associated individual. This automated process is crucial for reducing false alarms in crowded environments, ensuring that security personnel can focus on genuine threats.

The need for intelligent systems capable of distinguishing between benign and potentially hazardous situations has driven research in the field of computer vision and deep learning [1-3]. Early methods of AOD primarily relied on background subtraction techniques, which involved identifying static objects by analyzing changes in the scene over time. While effective in controlled environments, these methods often encountered limitations in real-world applications due to issues such as occlusion, sudden illumination changes, and dynamic backgrounds [4,5]. For instance, objects blending with the background or stationary individuals could trigger false alarms, reducing the overall efficiency of these systems. To mitigate these issues, adaptive dual-background models have been introduced, offering improved robustness by dynamically adjusting to scene changes. Additionally, techniques such as pixel-based finite-state machines (PFSM) have enhanced the ability to detect static objects [6]. The advent of deep learning has revolutionized AOD by enabling the development of sophisticated models that leverage convolutional neural networks (CNNs) to improve accuracy and robustness. Single-stage detection models, such as the YOLO series, and two-stage models, like Faster R-CNN, have demonstrated significant potential in object detection tasks, including AOD. These models excel in processing large datasets, extracting intricate features, and adapting to various environmental conditions [4,7,8]. For example, hybrid approaches integrating YOLO with contextual analysis and temporal consistency modeling have achieved higher accuracy rates in

979-8-3315-2223-0/25/$31.00 ©2025 IEEE

detecting abandoned objects in complex scenarios [1,6].

Another significant advancement in AOD research is the incorporation of ownership identification. Studies have focused on associating abandoned objects with their owners by analyzing gait patterns, spatial proximity, and behavioral cues. Techniques such as kernel canonical correlation analysis and gait energy image generation have been employed to identify and track individuals in surveillance footage, thereby enhancing the system's ability to determine ownership and potential intent [5,7,9]. Datasets such as PETS 2006, AVSS 2007, and ABODA [1] have played a crucial role in evaluating the performance of AOD systems. These datasets present diverse and challenging scenarios, including crowded environments, occlusions, and varying lighting conditions, making them invaluable for benchmarking and fine-tuning AOD frameworks [5,7]. Additionally, newly developed datasets have introduced complex cases, such as partially visible objects and group dynamics, to test the robustness of proposed systems [5,6]. Despite these advancements, challenges persist in detecting small or occluded objects, minimizing false positives, and achieving real-time performance. For instance, small objects often lack distinctive features, making them difficult to detect, particularly in low-resolution videos. To address these gaps, recent studies have proposed integrating advanced feature extraction methods, such as multi-scale fusion, with deep learning models tailored for small object detection [6,10].

This paper aims to contribute to the field by proposing an integrated AOD framework that combines real-time object tracking and deep feature analysis with Grounding DINO and fine-grained prompts. By leveraging state-of-the-art methodologies, the framework seeks to enhance detection accuracy, minimize false positives, and ensure scalability for real-world applications. The inclusion of Grounding DINO empowers the system to detect both expected and unforeseen objects, regardless of size, further bridging the gap between current limitations and the requirements of real-world scenarios. The proposed approach not only addresses technical challenges but also aims to improve public safety in diverse environments. The paper is organized into four sections to provide a comprehensive understanding of the study. Section I introduces the topic, highlighting the significance of abandoned object detection (AOD) and the proposed framework. Section II details the methodology, including Grounding DINO. Section III presents experimental results and analysis, while Section IV concludes with findings and future directions.

II. Methodology

The proposed methodology for detecting abandoned bags in video footage is outlined in the workflow below, as shown in Fig. 2. The approach integrates object detection, temporal frame sampling, and spatial association of detected objects using the Grounding DINO model. The system ensures robust detection of abandoned bags by tracking their association with individuals over time and checking for specific criteria based on distance and frame count. Below are the main steps involved in the process:

A. Input Videos and Frame Sampling

To efficiently process video data for abandoned bag detection, a frame sampling method was employed to reduce computational load while maintaining temporal information. The input video used in this research is derived from the ABODA footage dataset, which provides a diverse range of scenarios for abandoned bag detection. The total number of frames (N) and the frame rate (frames per second, FPS) were extracted using the OpenCV library, and these values were used to calculate a frame sampling interval (S). The sampling interval was determined using the formula:

$$S = max\left(\left\lfloor\frac{N}{T}\right\rfloor, 1\right) \quad (1)$$

where T represents the target number of frames to be processed (e.g., 500). This approach ensures that the sampled frames are evenly distributed across the video timeline while maintaining sufficient coverage of its temporal domain. Frames were then sampled at indices:

$$F_i = i \cdot S, i = 0,1,2, \dots, T-1 \quad (2)$$

and the selected frames were saved as images for further processing. By leveraging this method, the system balances accuracy and computational efficiency, enabling robust detection without the need to process every frame in the video. This systematic reduction of frames ensures that the workflow remains computationally feasible while preserving

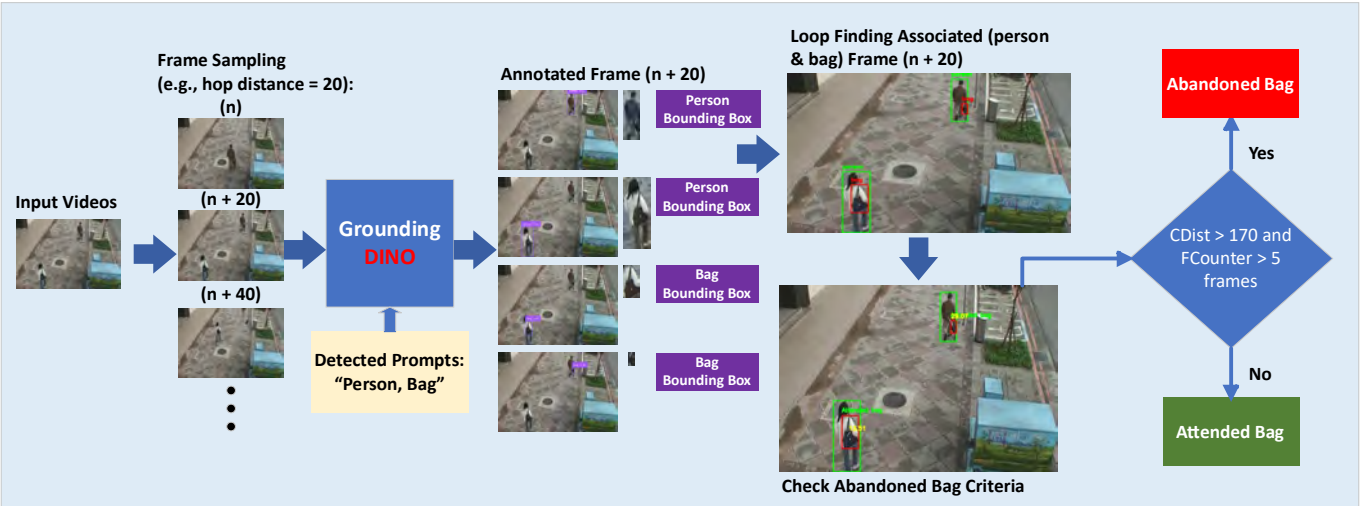

Fig. 2. Proposed workflow for Abandoned Object Detection using Frame Sampling (e.g., hop distance = 20). The process involves sampling frames from input videos, applying Grounding DINO to detect and annotate objects (e.g., person and bag), associating objects across frames through a looping mechanism, and evaluating abandoned bag criteria based on the centroid distance (CDist > 170) and the number of frames (FCounter > 5). The system distinguishes between an abandoned bag and an attended bag based on these parameters.

critical temporal details necessary for object detection and association.

B. Object Detection with Grounding DINO

The object detection phase leverages the Grounding DINO model, a state-of-the-art vision-language model designed to detect and label objects based on textual prompts. This model was configured to recognize specific objects such as "person" and "bag" using the defined text prompt P ("person, bag"). Each sampled frame, extracted during the frame sampling process, is processed by the Grounding DINO model to identify objects of interest. The model predicts bounding box coordinates (B), object confidence scores (C), and corresponding labels (L) for each detected object based on the input image (I) and prompt. This detection process can be mathematically expressed as:

$$(B, L, C) = \text{predict}(I, P, \text{box\_threshold}, \text{text\_threshold}) \quad (3)$$

where B represents the predicted bounding boxes defined by center coordinates (x_c,y_c), width (w), and height (h); C denotes the confidence scores for each detected object; and L corresponds to the object labels such as "person" or "bag." Detection thresholds for bounding box confidence and text relevance, defined as box_threshold and text_threshold, are used to filter out low-confidence predictions. In this research, we use box_threshold and text_threshold were set to 0.35 and 0.35, respectively. For each detected object, bounding box coordinates are transformed into pixel dimensions to accurately locate objects within the frame. This transformation ensures that the bounding box fits within the image boundaries. The bounding box parameters, including minimum ($x_{\min}$, $y_{\min}$) and maximum ($x_{\max}$, $y_{\max}$) pixel coordinates, are calculated as follows:

$$x_{min} = \max\left(0, \left\lfloor (x_C - \frac{w}{2}).W \right\rfloor\right) \quad (4)$$

$$y_{min} = \max\left(0, \left\lfloor (y_C - \frac{h}{2}).H \right\rfloor\right) \quad (5)$$

$$x_{max} = \min\left(W, \left\lfloor (x_C + \frac{w}{2}).W \right\rfloor\right) \quad (6)$$

$$y_{max} = \min\left(H, \left\lfloor (y_C + \frac{h}{2}).H \right\rfloor\right) \quad (7)$$

where W and H represent the width and height of the image, respectively. These calculations ensure the bounding boxes remain within the image bounds and accurately capture the detected objects.

To facilitate further analysis, the detected objects are annotated directly onto the frames using their bounding boxes and labels. Each annotated frame is saved for documentation, while the bounding box coordinates and associated labels are recorded in text files. Additionally, cropped regions corresponding to the bounding boxes are extracted and stored separately. This step allows the isolated analysis of specific objects, such as bags, in subsequent stages of the workflow. The object detection phase efficiently processes all sampled frames. In cases where no objects are detected, the system logs the absence of bounding boxes to maintain transparency in the detection process. This integration of the Grounding DINO model ensures robust detection and labeling of objects, forming the foundation for subsequent analysis of object relationships and behaviors. By combining vision-language capabilities with precise bounding box calculations, this methodology achieves high reliability in detecting and annotating objects, such as persons and bags, critical for abandoned bag detection.

C. Loop Finding for Object Association

The process of loop finding for object association is critical to reliably linking detected persons with their corresponding bags in consecutive video frames. This step ensures that objects are tracked consistently over time, minimizing the chances of false associations or missed detections. The primary goal is to associate each detected "person" bounding box with the closest "bag" bounding box within the same frame, using spatial proximity as the guiding criterion. To achieve this, the system calculates the centroids of bounding boxes, evaluates the Euclidean distances between centroids, and assigns unique identifiers to maintain consistent tracking of individuals and their associated objects. To determine the central point of a bounding box, the centroid is calculated as the average of the top-left and bottom-right coordinates of the box. For a given bounding box $b = (x_1, y_1, x_2, y_2)$, the centroid coordinates (c_x, c_y) are computed using the formula:

$$c_x = \frac{x_1 + x_2}{2}, \; c_y = \frac{y_1 + y_2}{2} \quad (8)$$

This centroid serves as the reference point for evaluating the spatial proximity between objects. The Euclidean distance between two centroids, such as those of a person and a bag, is calculated to quantify their spatial closeness. For two centroids (c_{x1}, c_{y1}) and (c_{x2}, c_{y2}), the distance d is determined by:

$$d = \sqrt{(c_{x2} - c_{x1})^2 + (c_{y2} - c_{y1})^2} \quad (9)$$

The bag with the smallest distance is assigned to the person, provided it has not already been associated with another individual in the same frame. This process is repeated for all detected persons, ensuring that every person is paired with the closest bag when possible. Unique identifiers are assigned to each person and their associated bag to enable consistent tracking across frames. To maintain temporal consistency in object tracking, a loop-checking mechanism is introduced. This mechanism monitors the associations between objects across consecutive frames. If a person and a bag remain unassociated for a predefined number of frames, the system flags the bag as "abandoned." This is achieved by maintaining a separation frame counter for each person-bag pair. When the separation counter exceeds a threshold T_s, the bag is marked as abandoned:

$$\text{SeparationFrames}(p, b) > T_S \Longrightarrow \text{AbandonedBag} \quad (10)$$

where p and b represent the person and bag bounding boxes, respectively. This mechanism ensures that objects are not prematurely flagged as abandoned due to temporary occlusions or brief separations. In addition to tracking, the system provides a visual representation of the detected associations by annotating each frame. Persons and bags are highlighted with bounding boxes, and lines are drawn to indicate associations. Labels are added to specify whether a bag is "attended" or "abandoned," enhancing the interpretability of the detection results. This approach effectively captures the dynamic interactions between individuals and their belongings, ensuring robust and accurate detection of abandoned bags. By leveraging spatial and temporal relationships, the system reduces false detections

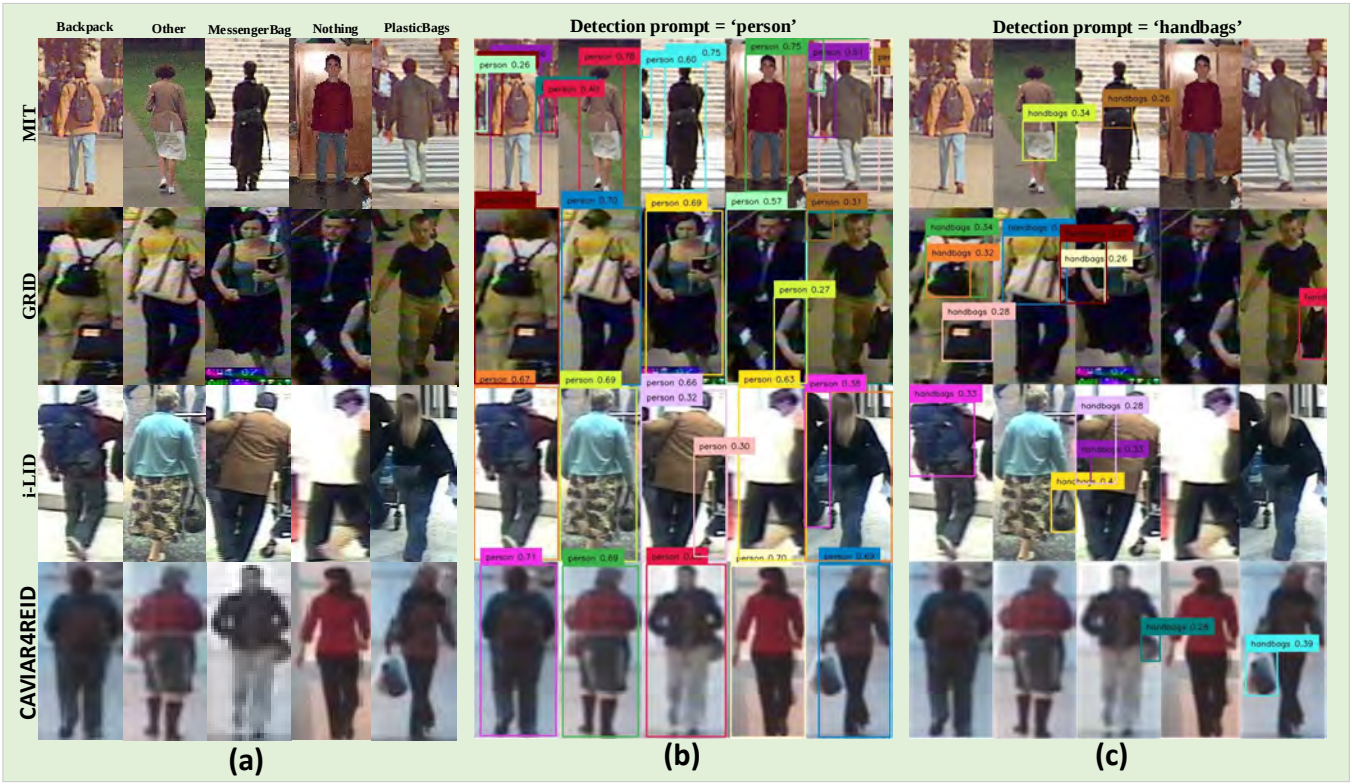

Fig. 3. (a) Examples from the PETA dataset showcasing various carrying attributes. (b) Detected results using the 'person' prompt. (c) Detected results using the 'handbags' prompt, highlighting specific carrying categories.

and enhances the reliability of real-time surveillance applications.

D. Abandoned Bag Criteria Check

The abandoned bag detection workflow processes each frame iteratively, applying the above criteria to all detected person-bag pairs. For instance, the abandonment criteria are defined as $T_s > 5$ frames and $C_{\text{dist}} > 170$ pixels, where T_s represents the separation threshold in frames and C_{dist} is the calculated distance between the person and the bag. If a pair meets these conditions for abandonment, the system generates alerts and saves annotated frames for review. This systematic approach ensures high accuracy in identifying abandoned bags by combining spatial and temporal analysis. By integrating distance thresholds, temporal tracking, and comprehensive annotations, this methodology achieves reliable abandoned bag detection in dynamic environments. The use of both spatial and temporal parameters minimizes false positives and enhances the overall robustness of the detection system.

III. Results and Discussion

The testing results of the proposed workflow for abandoned bag detection are analyzed. First, the performance of Grounding DINO is verified using the PETA [13] dataset to evaluate its ability to detect "person" and "bag" based on the prompts. Next, the workflow is tested using ABODA videos and, finally, compared with other research works.

A. Verification of Grounding DINO on the PETA Dataset

The PEdesTrian Attribute (PETA) dataset is a widely used benchmark for recognizing pedestrian attributes, including gender, clothing style, and carried objects, from images captured at significant distances. This dataset is particularly valuable in video surveillance scenarios where close-up shots of faces and bodies are often unavailable. It comprises 19,000 specifically the MIT, GRID, i-LID, and CAVIAR4REID subsets, to align with the research objectives. The PETA dataset provides images along with corresponding attribute labels, enabling precise analysis of specific features. In this research, the primary focus is on detecting persons and their attributes related to carrying bags. The dataset includes 11 carrying labels that are relevant to the "bag" category. These labels are: carrying BabyBuggy, carrying Backpack, carrying Other, carrying ShoppingTro, carrying Umbrella, carrying Folder, carrying LuggageCase, carrying MessengerBag, carrying Nothing, carrying PlasticBags, and carrying Suitcase. These labels cover a wide range of scenarios involving individuals carrying various types of bags and objects, ensuring the generalizability of the detection framework. The results of detecting attributes in the PETA dataset are summarized in Tables I and II. Table I highlights the detection performance for identifying persons, while Table II focuses on detecting the "bag" attribute across 11 carrying categories. These findings demonstrate the model's ability to accurately detect individuals and their associated carrying attributes, showcasing its potential for surveillance applications.

Figure 3 highlights five common carrying labels, showcasing diverse individual-bag scenarios within the dataset. Grounding DINO, using the "person" prompt on the PETA dataset, delivered exceptional results: 100% accuracy, recall, precision, and F1-score on the MIT subset; 98.43% accuracy and 99.21% F1-score on GRID; and near-perfect F1-scores on i-LID (99.79%) and CAVIAR4REID (99.96%), demonstrating high reliability in person detection. In contrast, Table II reveals challenges with the "bag" prompt. On the MIT subset, performance dropped to 65.00% accuracy, 61.00% recall, 55.00% precision, and 58.00% F1-score. The GRID subset performed worse, with 50.20% accuracy, 38.22% recall, 64.18% precision, and an F1-score of 47.91%, indicating inconsistency in bag detection.

TABLE I. PERFORMANCE OF GROUNDING DINO WHEN USING DETECTED PROMPTS "PERSON" ON PETA DATASET

Dataset	Subset	Tested Images	Prompts	TP	FP	TN	FN	Accuracy	Recall	Precision	F1-score
PETA [13]	MIT	888	'person'	888	0	0	0	100%	100%	100%	100%
	GRID	1275	'person'	1255	0	0	20	98.43%	98.43%	100%	99.21%
	i-LID	477	'person'	475	0	0	2	99.58%	100%	99.58%	99.79%
	CAVIAR4REID	1230	'person'	1229	0	0	1	99.92%	100%	99.92%	99.96%

TABLE II. PERFORMANCE OF GROUNDING DINO WHEN USING DETECTED PROMPTS "BAG" ON PETA DATASET

Dataset	Subset	Tested Images	Prompts	TP	FP	TN	FN	Accuracy	Recall	Precision	F1-score
PETA [13]	MIT	888	'bag'	213	174	364	137	65.00%	61.00%	55.00%	58.00%
	GRID	1275	'bag'	292	163	348	472	50.20%	38.22%	64.18%	47.91%
	i-LID	477	'bag'	135	110	139	93	57.44%	59.21%	55.10%	57.08%
	CAVIAR4REID	1230	'bag'	116	439	581	94	56.67%	55.24%	20.90%	30.33%

TABLE III. PERFORMANCE COMPARISON OF ABANDONED OBJECT DETECTION FOR ABODA DATASET WITH STATE-OF-THE-ART METHODS

ABODA Dataset	Ground Truth	Proposed		Newlin et al. [7]		Park et al. [11]		Lin et al. [1]		Dwivedi et al. [12]	
		TP	*FP*	*TP*	*FP*	*TP*	*FP*	*TP*	*FP*	*TP*	*FP*
video1	1	1	0	1	0	1	0	1	0	1	0
video2	1	1	0	1	0	1	0	1	0	1	0
video3	1	1	0	1	0	1	0	1	0	1	0
video4	1	1	0	1	0	1	0	1	0	1	0
video5	1	1	1	1	0	1	0	1	1	1	1
video6	2	1	1	2	1	2	0	2	0	1	3
video7	1	1	0	1	0	1	0	1	1	0	4
video8	1	1	0	1	0	1	0	1	1	0	3
video9	1	1	0	1	0	1	0	1	0	1	0
video10	1	1	0	1	0	1	0	1	0	1	0

The i-LID subset had an accuracy of 57.44%, recall of 59.21%, precision of 55.10%, and F1-score of 57.08%. Lastly, the CAVIAR4REID subset exhibited the lowest performance, with an accuracy of 56.67%, recall of 55.24%, precision of 20.90%, and F1-score of 30.33%. This poor performance in Table II highlights the difficulties in detecting "bag" due to several factors. First, some images are captured at significant distances from the camera, causing both the person and the bag to appear small in the frame, making detection harder. Additionally, certain ground truth annotations, such as "carryingOther," visually resemble bags, while others, like "carryingPlasticbags," are not actual bags, leading to misclassifications.

B. Tesing the Proposed Workflow on ABODA Videos

The proposed workflow for abandoned bag detection was evaluated using the ABODA dataset, which contains videos demonstrating various scenarios of individuals interacting with bags. The sequential detection process, as illustrated in Fig. 4, captures key stages in the abandonment workflow: (a) carrying the bag, (b) placing the bag, (c) walking away, and (d) identifying the bag as abandoned. This visual progression was tested across several videos (Videos 1, 3, and 9) to ensure robustness in detecting abandoned bags under diverse environmental conditions. Table III provides a detailed performance comparison between the ground truth and the proposed method. For each video, the ground truth represents the actual number of abandoned bag instances, while the performance of the proposed workflow is measured in terms of true positives (TP) and false positives (FP). The results reveal that the workflow accurately identified abandoned bags in most test videos, achieving a high true positive rate.

However, in Videos 5 and 6, the method encountered challenges, with some missed detections attributed to the specific environmental conditions. The main reason for missed detections in these videos was the use of night vision under low-light conditions, which caused the bags to appear either as glare or to blend in with the surrounding environment. This visual similarity between the bag and the background made it difficult for the workflow to accurately identify the object as an abandoned bag. Such challenges highlight the increased likelihood of false negatives in low-light conditions, where the model incorrectly predicts the absence of a bag. Additionally, the detection process is evaluated using key metrics:

1) True Positive (TP): The model correctly predicted the presence of a bag (e.g., a bag was correctly detected as a bag).

Fig. 4 Sequential detection of abandoned bags: (a) Carrying the bag, (b) Placing the bag, (c) Walking away, and (d) Bag identified as abandoned. Shown for Videos 1, 3, and 9.

TABLE IV. PERFORMANCE METRICS WITH GROUND TRUTH

Methods	Precision	Recall	F1-Score
Proposed	83.33%	90.91%	86.96%
Newlin [7]	91.67%	100%	95.65%
Park [11]	100%	100%	100%
Lin [1]	75.00%	81.82%	78.26%
Dwivedi [12]	47.62%	90.91%	62.50%

2) False Positive (FP): The model incorrectly predicted a bag where there was none (e.g., something that is not a bag was identified as a bag).

Despite these challenges, the workflow performed effectively in well-lit scenarios, as evidenced by the high accuracy and low false-positive rates in most videos. These findings suggest that improving the detection algorithm to better handle low-light environments, such as by integrating advanced preprocessing techniques for glare reduction or low-light enhancement, could further enhance its robustness and reliability. Future work should focus on refining the model to minimize false negatives and false positives, ensuring consistent performance across diverse lighting conditions and environmental challenges.

C. Comparison with Other Research Works

The proposed workflow for abandoned bag detection was compared with several state-of-the-art methods, including those by Newlin et al., Park et al., Lin et al., and Dwivedi et al., using the ABODA dataset. The comparison highlights the strengths and limitations of the proposed approach relative to existing methodologies, focusing on its performance across different scenarios captured in the dataset. As shown in Table III, the proposed method demonstrated strong performance, correctly identifying abandoned bags in most videos with minimal false positives. However, some challenges were observed in Videos 5 and 6, where false positives occurred. These errors were primarily due to low-light conditions and night vision challenges, where the bags either appeared as glare or blended into the background, making detection more difficult. Despite these challenges, the proposed method performed competitively, achieving a balance between true positives and false positives. In Table IV, the performance metrics—precision, recall, and F1-score—are derived based on true positives (TP) and false positives (FP) only, as the dataset focuses solely on abandoned bag detection. Since there is no explicit consideration of true negatives (TN) and false negatives (FN) in this evaluation, the metrics are calculated as follows:

1) Precision is calculated as $Precision = \frac{TP}{TP+FP}$, representing the proportion of correctly detected bags among all detected instances.

2) Recall is calculated as $\text{Recall} = \frac{TP}{Ground\ Truth\ Instances}$, representing the proportion of correctly detected bags relative to the total number of actual abandoned bags in the dataset.

3) F1-score is the harmonic mean of precision and recall, given by $F1 - score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

The proposed workflow achieved a precision of 83.33%, recall of 90.91%, and an F1-score of 86.96%, providing competitive performance despite slightly trailing top methods. Park et al.'s approach excelled with perfect metrics, while Newlin et al.'s method maintained high recall and F1-scores but lower precision. Lin et al. and Dwivedi et al. showed limitations with higher false-positive rates. Enhancements, such as advanced low-light image processing and refined feature extraction, could improve precision and reduce errors, ensuring consistent, accurate abandoned bag detection across diverse and challenging conditions.

IV. CONCLUSION

The proposed methodology integrates object detection, temporal frame sampling, and spatial association using the Grounding DINO model for abandoned bag detection in video footage. It leverages the PETA dataset for verification and the ABODA dataset for real-world testing. Frame sampling and bounding box transformations ensure computational efficiency without compromising accuracy, while centroid calculations and temporal tracking enhance reliability in person-bag pair detection. PETA results demonstrated robustness in detecting "persons" but revealed challenges with "bags" due to object size and annotation ambiguity. ABODA testing showed high detection rates in well-lit scenarios but limitations in low-light conditions, where glare and blending caused errors. The workflow achieved a competitive F1-score of 86.96%. Future improvements include addressing low-light challenges with advanced enhancement techniques to ensure consistent, reliable detection in diverse environments for real-time security applications.

REFERENCES

[1] K. Lin et al., "Abandoned Object Detection via Temporal Consistency Modeling and Back-Tracing Verification for Visual Surveillance," in *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1359-1370, July 2015, doi: 10.1109/TIFS.2015.2408263.

[2] M. T. Ahammed, S. Ghosh and M. A. R. Ashik, "Human and Object Detection using Machine Learning Algorithm," *2022 Trends in Electrical, Electronics, Computer Engineering Conference (TEECCON)*, Bengaluru, India, 2022, pp. 39-44.

[3] Smith, K., P. Quelhas, and D. Gatica-perez. "Detecting abandoned luggage items in a public space," In in PETS, 2006, pp. 75-82.

[4] A. M. Qasim, N. Abbas, A. Ali and B. A. A. -R. Al-Ghamdi, "Abandoned Object Detection and Classification Using Deep Embedded Vision," in *IEEE Access*, vol. 12, pp. 35539-35551, 2024.

[5] F. Amin, A. Mondal and J. Mathew, "A Large Dataset With a New Framework for Abandoned Object Detection in Complex Scenarios," in *IEEE MultiMedia*, vol. 28, no. 3, pp. 75-87, 1 July-Sept. 2021.

[6] Zhou, Lei, and Jingke Xu, "Enhanced Abandoned Object Detection through Adaptive Dual-Background Modeling and SAO-YOLO Integration" *Sensors* 24, no. 20: 6572, 2024.

[7] Russel, N.S. and Selvaraj, A., 2024. "Ownership of abandoned object detection by integrating carried object recognition and context sensing," *The Visual Computer*, *40*(6), pp.4401-4426.

[8] A. J. Amado-Garfias, S. E. Conant-Pablos, J. C. Ortiz-Bayliss and H. Terashima-Marín, "Improving Armed People Detection on Video Surveillance Through Heuristics and Machine Learning Models," in *IEEE Access*, vol. 12, pp. 111818-111831, 2024.

[9] D. Kim et al., "HLDNet: Abandoned Object Detection Using Hand Luggage Detection Network," in *IEEE Consumer Electronics Magazine*, vol. 11, no. 4, pp. 45-56, 1 July 2022.

[10] J. Kumar, S. Agarwal and J. S. Kumar, "YOLO Based Model for Garbage Detection," *2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, BHOPAL, India, 2022, pp. 1-5.

[11] Park, H., Park, S., Joo,Y., "Robustdetectionofabandonedobjectfor smart video surveillance in illumination changes," *Sensors* 19(23), 5114, 2019, https://doi.org/10.3390/s19235114.

[12] Dwivedi, N., Singh, D.K., Kushwaha, D.S.: An approach for unattended object detection through contour formation using back ground subtraction. Proc. Comput. Sci. 171, 1979–1988, 2020.

[13] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. "Pedestrian attribute recognition at far distance," *In Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792.