

# A Hybrid YOLO–Swin Transformer with FSM for Waste Disposal Behavior Analysis

Darniati<sup>1</sup>, Desi Anggreani<sup>2\*</sup>, Lukman<sup>3</sup>, Hasri Akbar Awal Rozaq<sup>4</sup>, A. Inayah Auliyah<sup>5</sup>, Bunga Dewi<sup>6</sup>,  
Muh. Ilham Akbar<sup>7</sup>

**Abstract**—The increasing volume of waste and improper disposal practices remain significant challenges in smart campus environments, despite the availability of waste management facilities. In this study, waste disposal behavior is defined as the temporal and contextual interaction among individuals, waste objects, and trash bins, encompassing both proper disposal and littering activities. To address this problem, this research proposes a hybrid deep learning framework that integrates YOLOv8 for real-time object detection, a finite state machine (FSM) for modeling object-state transitions during disposal activities, and a Swin Transformer for contextual behavior analysis. The proposed framework extends conventional object detection by combining spatial detection, temporal interaction modeling, and contextual reasoning to classify waste disposal behaviors from surveillance data. Furthermore, the integration of YOLOv8, FSM, and Swin Transformer provides a unified and interpretable architecture capable of capturing object interactions and behavioral patterns in complex surveillance environments. Experimental results indicate that the proposed method achieves a behavior classification mAP@50 of 93.8%, outperforming the FSM-only baseline (90.0%) and improving precision, recall, and F1 scores from 0.89, 0.84, and 0.86 to 0.92, 0.88, and 0.90, respectively. These findings demonstrate that the combination of spatial, temporal, and contextual analysis enables more accurate and reliable classification of waste disposal behavior, providing a practical solution for intelligent waste monitoring and sustainable smart campus management.

**Index Terms**—Hybrid deep learning, YOLOv8, finite state machine (FSM), swin transformer, behavior classification, littering detection.

Received: 4 April 2026; Revised: 16 June 2026; Accepted: 3 July 2026.

\*Corresponding author

<sup>1</sup>Darniati, Departement of Informatics, Universitas Muhammadiyah Makassar, Indonesia (e-mail: [darniati@unismuh.ac.id](mailto:darniati@unismuh.ac.id)).

<sup>2</sup>Desi Anggreani, Departement of Informatics, Universitas Muhammadiyah Makassar, Indonesia (e-mail: [desianggreani@unismuh.ac.id](mailto:desianggreani@unismuh.ac.id)).

<sup>3</sup>Lukman, Departement of Informatics, Universitas Muhammadiyah Makassar, Indonesia (e-mail: [lukman@unismuh.ac.id](mailto:lukman@unismuh.ac.id)).

<sup>4</sup>Hasri Akbar Awal Rozaq, Department of Computer Science, Gazi University, Türkiye (e-mail: [hakbar.rozaq@gazi.edu.tr](mailto:hakbar.rozaq@gazi.edu.tr)).

<sup>5</sup>A. Inayah Auliyah, Department of Information Systems, Institut Teknologi Bacharuddin Jusuf Habibie, Pare-Pare, Indonesia (e-mail: [inayah@ith.ac.id](mailto:inayah@ith.ac.id)).

<sup>6</sup>Bunga Dewi, Departement of Informatics, Universitas Muhammadiyah Makassar, Indonesia (e-mail: [105841121424@student.unismuh.ac.id](mailto:105841121424@student.unismuh.ac.id)).

<sup>7</sup>Muh. Ilham Akbar, Departement of Informatics, Universitas Muhammadiyah Makassar, Indonesia (e-mail: [105841105822@student.unismuh.ac.id](mailto:105841105822@student.unismuh.ac.id)).

## I. INTRODUCTION

University campuses represent complex micro-environments where academic, residential, and commercial activities continuously generate significant amounts of waste. Despite the implementation of recycling programs and waste separation facilities, improper waste disposal behavior among campus users remains a persistent issue [1]. Examples of such behavior include littering near disposal areas, leaving waste beside trash bins instead of placing it inside, or abandoning waste in public spaces. These actions contribute to environmental degradation, increase operational costs, and reduce the effectiveness of campus sustainability programs. Consequently, intelligent monitoring systems capable of automatically identifying and analyzing waste disposal behavior are increasingly needed to support sustainable smart campus management strategies [2].

Recent advances in artificial intelligence, particularly in computer vision and deep learning, have enabled the development of automated environmental monitoring systems capable of processing surveillance data in real time [3]. Existing studies have widely applied deep learning methods for waste classification, recycling automation, and object detection. Most of these approaches focus primarily on identifying waste categories, detecting trash bins, or recognizing recyclable materials using static object-level recognition techniques [4]. Although these methods can accurately detect objects, they are not designed to understand behavior. Object detection focuses on identifying the presence, category, and location of objects within an image, whereas behavior analysis requires reasoning about temporal interactions, contextual relationships, and object-state transitions across multiple frames [5], [6]. For instance, identifying littering behavior requires determining whether a person releases a waste object outside a trash bin area and whether the object remains in the scene afterward. Such decisions depend on sequential interactions between people, waste objects, and environmental context, which cannot be inferred reliably from frame-based object recognition alone.

YOLO-based object detection frameworks have become widely adopted in real-time surveillance applications due to their high detection accuracy and computational efficiency [7]. Several studies have demonstrated the effectiveness of

YOLO-based approaches in environmental monitoring, waste detection, and smart city surveillance systems [8]. YOLO is particularly suitable for surveillance environments because it performs object localization and classification in a single-stage architecture, enabling real-time operation with low computational overhead.

Meanwhile, transformer-based architectures have demonstrated strong capability in modeling global contextual relationships through self-attention mechanisms. Unlike conventional convolutional neural network (CNN)-based methods that mainly focus on local spatial features, transformer-based models are capable of capturing broader contextual dependencies between objects within an image [9]. Among these approaches, the Swin Transformer has demonstrated strong capability in hierarchical contextual representation while maintaining computational scalability for high-resolution visual inputs [10]. Recent hybrid CNN–Transformer approaches have further shown promising improvements in contextual feature extraction and object representation in complex visual environments. Moreover, several studies integrating transformer modules into CNN-based detection architectures have reported enhanced performance in challenging object recognition scenarios [11].

Despite these advances, existing YOLO-based and hybrid YOLO–Transformer approaches remain primarily oriented toward object recognition performance. Most studies focus on improving detection accuracy, feature extraction quality, or contextual representation, while providing limited support for temporal reasoning, behavioral semantics, and interpretable decision-making. As a result, current approaches are insufficient for behavior-centric surveillance applications that require understanding how objects interact over time rather than simply identifying their presence [12]. Furthermore, no existing study has specifically integrated YOLOv8, Swin Transformer, and Finite State Machine (FSM) within a unified framework for waste disposal behavior classification in smart campus environments. Previous studies largely remain limited to static waste detection or object classification and do not explicitly model the temporal interactions among individuals, waste objects, and disposal facilities that are essential for distinguishing proper disposal from littering behavior [13]. This represents a significant research gap because intelligent waste monitoring systems require not only accurate object detection but also contextual and temporal understanding of disposal activities [14]. Therefore, a more comprehensive framework is required that integrates spatial object detection, temporal interaction reasoning, and contextual analysis within a unified behavioral understanding system capable of operating in real-world surveillance environments.

To address this gap, this study proposes a hybrid behavior-analysis framework that integrates YOLOv8 for real-time object detection, Swin Transformer for contextual relationship modeling, and FSM for interpretable temporal state-transition reasoning. YOLOv8 detects relevant entities, including persons, waste objects, and trash bins, while FSM models interaction states during disposal activities and Swin

Transformer refines behavior interpretation through contextual analysis. The integration of these components enables the framework to combine spatial detection, temporal reasoning, and contextual understanding within a single architecture for waste disposal behavior classification.

The contributions of this study are threefold. First, it introduces a behavior-centric monitoring framework capable of automatically distinguishing proper disposal and littering behaviors from surveillance video streams. Second, it proposes a novel architectural integration of YOLOv8, Swin Transformer, and FSM that combines real-time object detection, contextual understanding, and interpretable temporal reasoning within a unified system. Third, it demonstrates that the combination of spatial, temporal, and contextual analysis improves the reliability of behavior classification for intelligent waste monitoring in complex smart campus environments.

## II. RELATED WORK

Recent studies have demonstrated the effectiveness of deep learning for waste classification and intelligent waste management systems. Existing approaches have successfully improved waste recognition, sorting efficiency, and recycling automation through image-based classification models and IoT-assisted monitoring systems [15], [16], [17]. These methods primarily focus on identifying waste categories and improving operational efficiency in waste management processes. However, they generally treat waste as isolated objects and do not consider the behavioral context surrounding disposal activities, such as how individuals interact with waste objects and disposal facilities over time.

YOLO-based architectures have become widely adopted in real-time surveillance applications because they provide an effective balance between detection accuracy and computational efficiency [18], [19]. Recent YOLO variants have achieved state-of-the-art performance in environmental monitoring, smart city applications, and waste detection tasks. Although these approaches can accurately detect persons, waste objects, and disposal facilities, their outputs are limited to object localization and classification. Consequently, they provide spatial information but do not explicitly model the temporal interactions required for behavioral interpretation.

Transformer architectures have significantly advanced contextual representation learning in computer vision. DETR introduced end-to-end transformer-based object detection, while Swin Transformer further improved contextual modeling through hierarchical feature representation and shifted-window self-attention mechanisms [20], [21]. Recent hybrid CNN–Transformer architectures have demonstrated improved robustness in complex visual environments by combining local feature extraction with global contextual reasoning [22], [23], [24]. Nevertheless, these approaches mainly focus on improving feature representation and recognition accuracy. They rarely incorporate explicit temporal reasoning mechanisms needed to interpret behavioral events and interaction sequences.

Finite State Machines (FSMs) have been widely applied in

domains requiring interpretable state-transition modeling, including traffic control systems, autonomous decision-making, simulation environments, and behavior-driven applications[25], [26]. FSM-based approaches provide transparent decision logic by representing behaviors as transitions between predefined states. In surveillance and activity-recognition contexts, FSMs have also been used indirectly to model event progression and action sequences. However, existing FSM implementations are typically rule-based and often operate independently of contextual visual understanding. As a result, their performance may degrade under ambiguous scenes, occlusion, or complex multi-object interactions. Furthermore, the application of FSM for waste disposal behavior analysis remains largely unexplored in the literature.

The reviewed literature indicates that existing studies address only partial aspects of waste disposal monitoring. Waste classification systems focus on waste categories, YOLO-based approaches provide efficient object detection, transformer-based models enhance contextual understanding, and FSMs offer interpretable temporal reasoning. However, no existing framework comprehensively integrates real-time object detection, contextual scene understanding, and interpretable temporal state-transition modeling for waste disposal behavior classification. More specifically, studies combining YOLO and Transformers generally target object recognition tasks, whereas FSM-based approaches rarely incorporate deep contextual visual representations. To address this gap, the proposed framework integrates YOLOv8, Swin Transformer, and FSM within a unified architecture. YOLOv8 provides real-time spatial detection, Swin Transformer captures contextual relationships among detected entities, and FSM models temporal state transitions during disposal activities. This integration enables the system to classify disposal behaviors, including proper disposal and littering, rather than merely detecting objects within the scene.

Table 1.

Comparison of Related Studies and the Proposed Framework

| Reference                 | Waste Classification | Object Detection | Context Modeling | Temporal Reasoning | Behavior Classification |
|---------------------------|----------------------|------------------|------------------|--------------------|-------------------------|
| [15]                      | ✓                    | X                | X                | X                  | X                       |
| [16]                      | ✓                    | X                | X                | X                  | X                       |
| [17]                      | ✓                    | X                | X                | X                  | X                       |
| [18], [19]                | X                    | ✓                | X                | X                  | X                       |
| [20], [21]                | X                    | ✓                | ✓                | X                  | X                       |
| [22], [23], [24]          | X                    | ✓                | ✓                | X                  | X                       |
| [25], [26]                | X                    | X                | X                | ✓                  | Partial                 |
| <b>Proposed Framework</b> | ✓                    | ✓                | ✓                | ✓                  | ✓                       |

The comparison highlights that existing studies generally address object recognition, contextual representation, or temporal reasoning independently. In contrast, the proposed framework integrates all these components within a unified

architecture for waste disposal behavior classification.

### III. RESEARCH METHOD

#### A. System Architecture

This study proposes a hybrid framework for waste disposal behavior classification in smart campus environments by integrating spatial object detection, temporal interaction modeling, and contextual scene understanding. Unlike conventional computer vision approaches that analyze objects independently in individual frames, the proposed framework models the complete waste disposal process by jointly analyzing interactions among persons, waste objects, and trash bins across sequential observations.

Figure 1 illustrates the overall processing pipeline of the proposed framework. The system begins with surveillance video acquisition, followed by frame extraction and object detection using YOLOv8. The detected objects are then associated across consecutive frames using an object tracking algorithm to preserve temporal identities. The resulting object trajectories are processed by a Finite State Machine (FSM) to infer behavioral state transitions, while the Swin Transformer extracts contextual features from the surrounding scene. Finally, the outputs from both modules are integrated to classify the observed waste disposal behavior as either proper disposal or littering. Let the input to the system be a video sequence  $V$  captured by a surveillance camera, represented as a sequence of  $T$  consecutive frames:

$$V = \{f_1, f_2, \dots, f_T\}, \quad f_t \in \mathbb{R}^{H \times W \times C} \quad (1)$$

where  $H$ ,  $W$ , and  $C$  denote the height, width, and number of channels (RGB) of each frame at time step  $t$ . The ultimate goal of the system is to map the sequence  $V$  to a behavior class  $Y \in y$ , where  $y = \{\text{Proper disposal, littering, no disposal activity}\}$ .

#### B. Datasets and Data Preparation

The dataset used in this study consists of surveillance data collected in campus waste disposal areas to capture real-world waste disposal activities. A total of 150 surveillance videos were recorded to capture interactions among three primary entities: persons, waste objects, and trash bins under realistic disposal scenarios. The recordings include both proper disposal and littering behaviors and were acquired under varying environmental conditions, including morning, afternoon, and evening illumination, multiple camera viewpoints, different object distances, partial occlusions, and complex backgrounds to improve model robustness and generalization.

The dataset also includes 200 manually annotated images, which are representative frames extracted from the 150 surveillance videos rather than independently collected images. These representative frames were selected to maximize the diversity of object appearances, interaction scenarios, and environmental conditions for object detector training and

validation. Therefore, the complete video sequences provide the temporal information required for behavior analysis, whereas the annotated image samples are used to train and evaluate the object detection model.

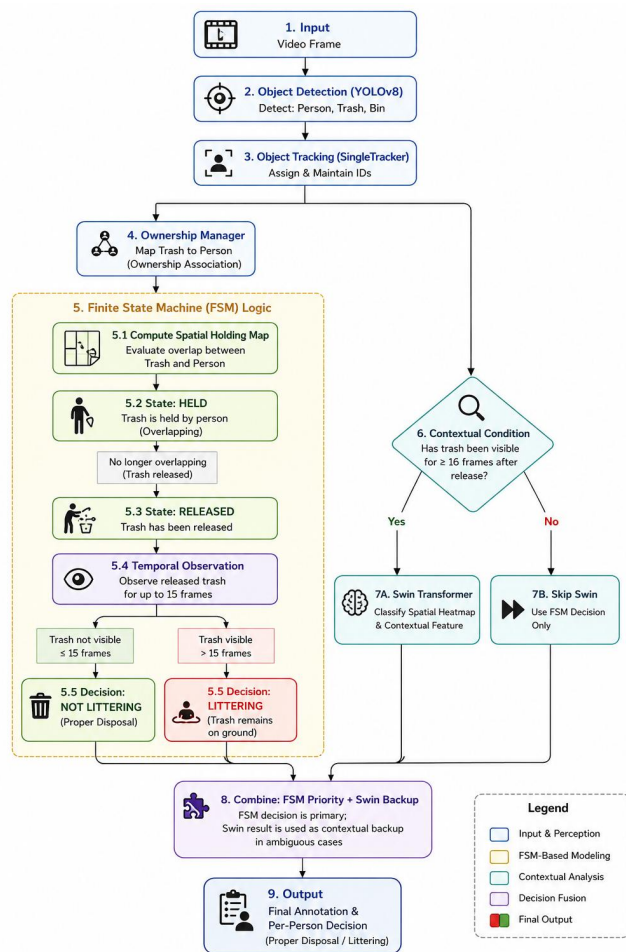


Fig. 1. Proposed hybrid framework for waste disposal behavior classification using YOLOv8, tracking, FSM, and swin transformer.

As illustrated in Figure 1, each surveillance video is first converted into sequential image frames before object detection and behavior analysis. Frames were extracted at a sampling rate of 15 frames per second, which was selected to preserve temporal continuity between consecutive disposal actions while reducing redundant frames with highly similar visual content. This frame extraction strategy provides sufficient temporal resolution for subsequent tracking and behavior modeling while maintaining computational efficiency.

Following frame extraction, representative frames were manually annotated using bounding-box labeling for three object categories: person, trash, and trash bin. Annotation was performed according to predefined labeling guidelines specifying object boundaries, labeling priorities under overlap conditions, and the treatment of partially visible objects. To improve annotation quality, all labeled samples underwent an additional verification process, and ambiguous annotations were reviewed and corrected before model development. Table

2 presents representative samples of the surveillance videos and annotated images used in this study for behavior analysis and object detection training.

Table 2.  
Characteristics of Data Sets

| Dataset                                                                            | Description                                         |
|------------------------------------------------------------------------------------|-----------------------------------------------------|
|  | Surveillance videos for temporal behavior analysis. |
|  | Image samples used for object detection training.   |

The final annotated dataset was divided into 70% training data, 20% validation data, and 10% testing data. The training subset was used to optimize the YOLOv8 detector, the validation subset was employed for hyperparameter tuning, and the testing subset was reserved exclusively for performance evaluation. Before model training, images were resized to  $640 \times 640$  pixels, normalized, and standardized to improve data consistency, stabilize the learning process, and enhance model generalization across different surveillance conditions. Meanwhile, the original video sequences were retained for the subsequent object tracking, FSM-based temporal reasoning, and contextual behavior analysis stages, enabling the proposed framework to exploit both spatial and temporal information throughout the classification process.

### C. YOLO-based Object Detection

You Only Look Once (YOLO) is a deep learning-based object detection algorithm that performs object localization and classification simultaneously within a single-stage detection architecture. Unlike conventional two-stage detectors that separately generate object proposals and classification results, YOLO processes the entire image in a single forward pass, enabling faster inference while maintaining competitive detection accuracy. As the latest generation of the YOLO family, YOLOv8 introduces improvements in feature extraction capability, localization precision, and computational efficiency, making it suitable for real-time surveillance and environmental monitoring applications [27].

In this study, YOLOv8 is employed to detect objects associated with waste disposal behavior, namely person, trash, and trash bin. The detection process begins by receiving extracted video frames as input, followed by feature extraction to capture visual characteristics and spatial information from the scene. The model then simultaneously predicts object positions and categories, producing three outputs for each detected object: bounding box coordinates, class labels, and confidence scores. These outputs provide structured spatial representations that describe object location and detection certainty within each frame.

The detection results generated by YOLOv8 serve as the initial perception layer in the proposed framework and provide essential information for subsequent analysis stages. However, object detection alone is insufficient for behavior understanding because recognizing object presence in individual frames cannot explain behavioral outcomes. Therefore, the detection outputs are further integrated with object tracking, FSM-based temporal modeling, and Swin Transformer contextual analysis to enable comprehensive classification of waste disposal behavior into proper disposal and littering categories.

Table 3.  
YOLOv8 Configuration Parameters

| Parameter            | Value                                       | Description                                       |
|----------------------|---------------------------------------------|---------------------------------------------------|
| Model                | YOLOv8                                      | Object detection architecture                     |
| Input size           | 640 × 640                                   | Input image resolution                            |
| Object classes       | 3                                           | Person, trash, trash bin                          |
| Batch size           | 16                                          | Number of samples per iteration                   |
| Epochs               | 100                                         | Total training iterations                         |
| Learning rate        | 0.001                                       | Initial learning rate                             |
| Optimizer            | AdamW                                       | Parameter optimization algorithm                  |
| Confidence threshold | 0.50                                        | Minimum confidence for valid detection            |
| Iou threshold        | 0.50                                        | Threshold for overlap filtering                   |
| Loss function        | Detection Loss (box, cls, dfl)              | Combined loss for localization and classification |
| Data split           | 70% / 20% / 10%                             | Training/validation/ testing                      |
| Output               | Bounding box, class label, confidence score | Detection results                                 |

Figure 2 illustrates the characteristics of the annotated dataset used for YOLOv8 training. The visualization presents the distribution of object instances across classes, spatial locations of bounding boxes, and object size variation. These distributions indicate the diversity of object appearance and placement within the dataset, supporting model learning under different environmental conditions and interaction scenarios.

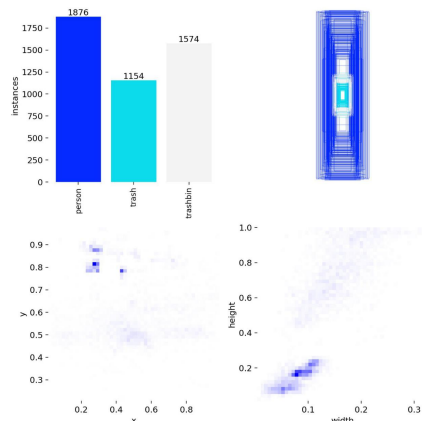


Fig. 2. Distribution of annotated objects and bounding box characteristics used for YOLOv8 training.

#### D. FSM-Based Behavior Modeling

This study employs a Finite State Machine (FSM) to model waste disposal behavior based on temporal interactions between detected objects. FSM is used because behavior cannot be determined from a single frame and requires observation of object-state changes over time[28]. Through predefined states and transitions, FSM enables interpretable behavior classification based on sequential interactions between person and trash objects. As illustrated in Figure 3, the FSM begins by evaluating the relationship between the detected person and trash object. When the trash remains associated with the person, the system enters S0 (HELD). Once the trash is released and no longer overlaps with the person, the FSM transitions to S1 (RELEASED). The system then observes the released object for 15 frames. If the trash is no longer visible, the behavior is classified as S2 (NOT LITTERING/Proper Disposal). Otherwise, if the trash remains visible beyond the observation period, the behavior transitions to S3 (LITTERING). This temporal modeling allows the system to convert object interactions into final disposal behavior decisions.

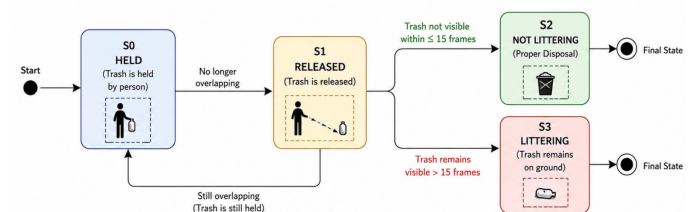


Fig. 3. Finite State Machine (FSM)-based behavior modeling for waste disposal classification.

#### E. Swin Transformer Context Modeling

To improve behavior interpretation under complex visual conditions, this study employs the Swin Transformer as a contextual analysis module. Unlike conventional convolution-based approaches that primarily capture local spatial features, the Swin Transformer utilizes hierarchical feature extraction and shifted-window self-attention to model contextual relationships among detected objects and their surrounding environment [29]. Within the proposed framework, the Swin Transformer complements the FSM by providing scene-level contextual reasoning when temporal state transitions alone are insufficient to determine the final disposal behavior.

In the proposed framework, the Swin Transformer is activated after the FSM-based temporal modeling stage whenever additional contextual verification is required. Specifically, contextual analysis is performed when the released waste object remains visible during the observation window or when the spatial relationship between the waste object and the trash bin is ambiguous. In such situations, temporal information alone cannot reliably determine whether the waste has been properly disposed of or intentionally discarded outside the disposal area. The contextual reasoning

process is performed using the self-attention mechanism:

$$\text{Attention}(Q, K, V = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

where ( $Q$ ), ( $K$ ), and ( $V$ ) denote the query, key, and value feature representations extracted from image patches. In the proposed framework, the query corresponds to the released waste object, whereas the key and value vectors encode contextual information from neighboring regions, including the detected person, trash bin, and surrounding environment. The self-attention mechanism evaluates the relationships among these visual features to emphasize context relevant to waste disposal behavior while suppressing irrelevant background information.

The contextual features generated by the Swin Transformer complement the temporal reasoning provided by the FSM, particularly in ambiguous situations where object position or temporal state transitions alone cannot reliably distinguish proper disposal from littering. By incorporating contextual relationships, the proposed framework can resolve cases involving partial occlusion, object proximity to the trash bin, or complex backgrounds. The contextual representation is subsequently fused with the FSM output to produce the final behavior classification, resulting in more robust and reliable littering detection under real-world surveillance conditions.

#### F. Evaluation Metrics

This study employs four evaluation metrics, namely precision, recall, F1-score, and mean Average Precision (mAP), to assess the performance of the proposed framework. Precision, Recall, and F1-score are used to evaluate the accuracy of waste disposal behavior classification, whereas mAP is employed to measure the object detection performance of the YOLOv8 detector. Together, these metrics provide a comprehensive evaluation of both the spatial detection capability and the overall behavior classification performance of the proposed framework.

Precision measures the proportion of correctly predicted positive instances among all positive predictions generated by the model and is defined as [30].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

where ( $TP$ ) denotes the number of true positive predictions and ( $FP$ ) represents the number of false positive predictions. Recall measures the ability of the model to identify all relevant positive instances in the dataset and is calculated as.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

where ( $FN$ ) denotes the number of false negative predictions. The harmonic mean of Precision and Recall is expressed by the F1-score,

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

which provides a balanced evaluation when both false positives and false negatives are considered.

Finally, the object detection performance is evaluated using mean Average Precision (mAP), which summarizes the area under the Precision–Recall curve across all object classes and confidence thresholds. A higher mAP value indicates better

localization and classification performance of the object detector, providing reliable spatial information for the subsequent tracking, FSM-based temporal reasoning, and contextual behavior classification stages.

## IV. RESULT

This section presents the experimental results of the proposed framework and evaluates the contribution of each processing stage. The process includes YOLOv8 for object detection, object tracking for temporal continuity, FSM for behavior state modeling, and Swin Transformer for contextual refinement. The results are analyzed to demonstrate how the integration of spatial, temporal, and contextual information improves waste disposal behavior classification.

### A. Object Detection Performance (YOLOv8)

The object detection stage evaluates the capability of YOLOv8 to identify objects involved in waste disposal activities, namely person, trash, and trash bin. Each extracted video frame is processed independently to generate object location and category information. The outputs produced at this stage become the initial spatial representation used for temporal and behavioral analysis in subsequent processes.

Figure 4 illustrates the detection process across sequential video frames. YOLOv8 processes each frame through several stages: (1) feature extraction to capture visual patterns from the input image, (2) object localization to estimate bounding box positions, (3) object classification to assign detected regions into predefined classes, and (4) confidence estimation to determine prediction certainty. This process allows the model to continuously detect relevant objects throughout the video sequence and maintain detection consistency under different object positions and scene conditions.

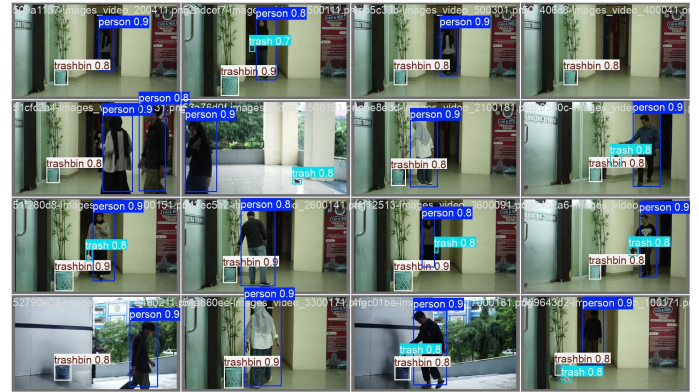


Fig. 4. Sequential object detection results generated by YOLOv8 across multiple video frames.

Figure 5 presents an example of the final detection output generated by YOLOv8 on a single frame. The model successfully identifies the detected objects and provides bounding boxes, class labels, and confidence scores as structured spatial information. These outputs serve as the input for the object tracking stage to preserve object continuity and support behavior classification in later stages.



Fig. 5. Object detection output generated by YOLOv8.

To evaluate the object identification capability of the proposed framework, YOLOv8 performance was analyzed using the Precision–Recall (PR) Curve. This evaluation measures the model’s ability to maintain detection accuracy and completeness across different confidence thresholds. Since object detection serves as the initial stage of behavior analysis, reliable detection is required to support tracking and subsequent behavior classification.

Figure 6 shows that YOLOv8 achieved high detection performance with an overall mAP@0.5 of 0.985 across three object classes: person, trash, and trash bin. The trash bin class obtained the highest performance (0.991), followed by person (0.991) and trash (0.974). The slightly lower result for the trash class indicates greater detection difficulty due to smaller object size and more varied object appearance.

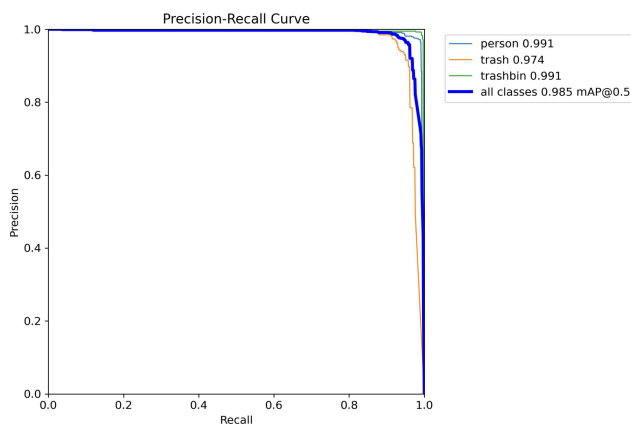


Fig. 6. Precision–Recall curve of YOLOv8 object detection performance for person, trash, and trash bin classes.

These results demonstrate that the proposed detection stage provides sufficiently accurate spatial information for subsequent analysis. High precision reduces incorrect object associations during tracking, while strong recall helps preserve interaction continuity, supporting more reliable temporal modeling and final waste disposal behavior classification.

To further evaluate the classification capability of YOLOv8, a normalized confusion matrix was used to analyze the prediction distribution across object classes. This evaluation

complements the precision–recall analysis by providing a more detailed interpretation of classification consistency and identifying potential confusion between detected objects and background regions.

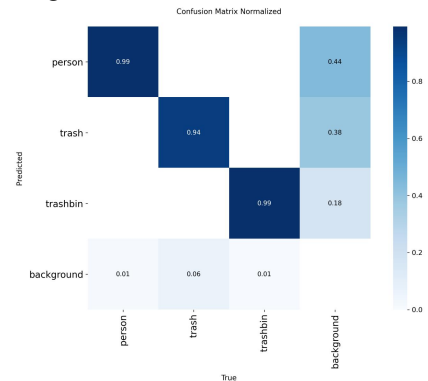


Fig. 7. Normalized confusion matrix of YOLOv8 object detection results for person, trash, trash bin, and background classes.

Figure 7 shows that most prediction values are concentrated along the diagonal area, indicating high classification consistency. The person and trash bin classes achieved the highest correct classification values of 0.99, while the trash class achieved 0.94. These results indicate that YOLOv8 successfully distinguished the three target objects across different observation conditions. However, several detections were still associated with the background class, with confusion values of 0.44 for person, 0.38 for trash, and 0.18 for trash bin. Meanwhile, background misclassification toward target objects remained low, with values of 0.01, 0.06, and 0.01 for person, trash, and trash bin, respectively.

These results indicate that the proposed detection stage provides reliable spatial information for subsequent processing stages. Nevertheless, the lower performance observed in the trash class suggests that smaller object dimensions and visual similarity to surrounding environments remain challenging conditions that may influence tracking stability and temporal behavior classification at optimal levels of precision under a wide range of scale and distance conditions.

### B. Tracking Stability and Temporal Representation

The tracking process was applied after the object detection stage to maintain object identity and represent interactions continuously across video frames. Unlike object detection that analyzes each frame independently, tracking associates detected objects over time by assigning a unique identifier (ID) and updating their positions in subsequent frames. This process enables the system to construct object trajectories and preserve relationships between person, trash, and trash bin throughout the observed activity.

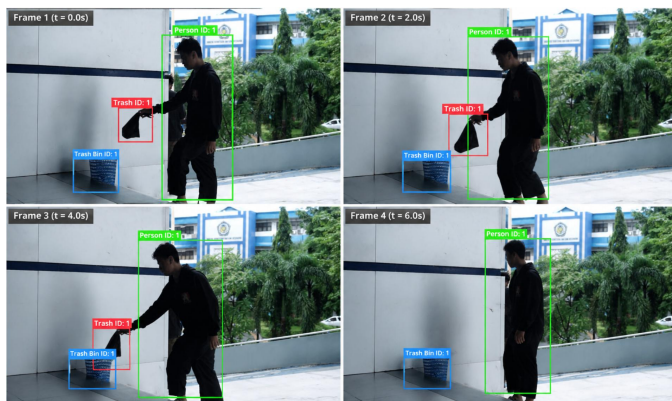


Fig. 8. Visual tracking across consecutive frames.

Figure 8 illustrates the tracking process across consecutive video frames. After YOLOv8 detects the relevant objects, the tracking module assigns a consistent identity to each object, including Person ID: 1, Trash ID: 1, and Trash Bin ID: 1. These identities are maintained from the first frame to the next frames, showing that the system can preserve object continuity during movement.

Step by step, the person is first detected while holding the trash object near the bin. In the following frames, the tracker maintains the same object IDs while the person and trash change position. This indicates that the tracking module successfully captures temporal movement and interaction between the person, trash, and trash bin rather than treating each frame independently.

The result shows that tracking stability is important for behavior classification because consistent object IDs allow the system to identify interaction changes over time. The generated temporal information is then used as input for FSM-based behavior modeling to determine the final disposal behavior.

### C. FSM-Based Behavior Classification Results

This subsection evaluates the effectiveness of the Finite State Machine (FSM) in transforming tracked object interactions into final behavioral decisions. Unlike object detection and tracking, which focus on spatial identification and temporal continuity, the FSM performs temporal reasoning by analyzing changes in the interaction between the detected person and waste object across sequential frames. The FSM determines object interaction based on the spatial relationship between the person and the waste object, which is computed using the Intersection over Union (IoU) between their corresponding bounding boxes. In this study, an interaction is classified as Overlapping when the IoU is greater than zero, indicating that the waste object is still associated with the detected person. Conversely, the interaction is classified as Separated when the IoU equals zero, indicating that the waste object is no longer in contact with the person. After entering the Released state, the FSM observes the tracked waste object for 15 consecutive frames before assigning the final behavioral decision.

Table 4.  
FSM state Transitions and Final Behavior Classification Across Consecutive Frames

| Frame | Person ID      | Trash ID       | Spatial Relationship (IoU)      | FSM State | Final Verdict |
|-------|----------------|----------------|---------------------------------|-----------|---------------|
| 240   | P <sub>1</sub> | T <sub>1</sub> | Overlapping (IoU > 0)           | Held      | Tracking      |
| 250   | P <sub>1</sub> | T <sub>1</sub> | Overlapping (IoU > 0)           | Held      | Tracking      |
| 300   | P <sub>1</sub> | T <sub>1</sub> | Overlapping (IoU > 0)           | Held      | Tracking      |
| 345   | P <sub>1</sub> | T <sub>1</sub> | Separated (IoU = 0)             | Released  | Tracking      |
| 360   | P <sub>1</sub> | T <sub>1</sub> | Separated (IoU = 0 > 15 frames) | Released  | Littering     |

Table 4 illustrates the state transitions generated by the FSM during the waste disposal process. At frames 240–300, the waste object remained spatially associated with the detected person (IoU > 0), indicating that the object was still being carried. Consequently, the FSM maintained the Held state and continued tracking the interaction. At frame 345, the IoU decreased to zero, indicating that the waste object had been released, which triggered the transition to the Released state and initiated the observation period. If the released waste object remains separated from the detected person for more than 15 consecutive frames, the FSM assigns the final behavior as Littering, as observed at frame 360. The 15-frame observation window was selected based on the video sampling rate of 10 fps, corresponding to approximately 1.5 seconds, which was found sufficient during preliminary experiments to distinguish temporary release during disposal from persistent littering behavior. These results demonstrate that the FSM successfully converts tracked object interactions into interpretable behavioral decisions by combining spatial relationships with temporal persistence rather than relying on isolated frame observations.

### D. Contribution of Swin Transformer for Contextual Understanding

This subsection evaluates the contribution of Swin Transformer in improving waste disposal behavior classification through contextual understanding. While the FSM stage performs temporal reasoning based on object interactions and state transitions, behavior interpretation may become uncertain when released objects remain visible, interactions are partially occluded, or environmental conditions introduce ambiguity. Therefore, Swin Transformer is incorporated as a contextual refinement mechanism to complement temporal decisions and improve final classification reliability.



Fig. 9. Contextual features generated by Swin Transformer for behavior classification.

Figure 9 illustrates the contextual representation generated by the Swin Transformer during behavior analysis. The highlighted regions indicate areas receiving higher attention during feature extraction, particularly around the detected person, trash, and surrounding interaction space. Unlike object detection and FSMs that focus on object presence and temporal transitions, Swin Transformer captures broader contextual relationships within the scene. This enables the model to evaluate whether the observed object placement corresponds to proper disposal or littering under complex environmental conditions.

Table 5.  
FSM State Transitions and Final Behavior Classification Across Consecutive Frames.

| Method                            | Precision | Recall | F1-Score | mAP   |
|-----------------------------------|-----------|--------|----------|-------|
| FSM Only                          | 0.89      | 0.84   | 0.86     | 0.90  |
| FSM + Swin Transformer (Proposed) | 0.92      | 0.88   | 0.90     | 0.93  |
| Improvement                       | +0.03     | +0.04  | +0.04    | +0.03 |

Table 5 presents the comparative performance between FSM-only and the proposed framework integrating FSM and Swin Transformer. The results show improvements across all evaluation metrics, where precision increased from 0.89 to 0.92, recall from 0.84 to 0.88, F1-score from 0.86 to 0.90, and mAP from 0.90 to 0.93. These findings indicate that contextual analysis improves the preservation of interaction information and reduces ambiguous predictions. Overall, the integration of Swin Transformer demonstrates that combining spatial detection, temporal reasoning, and contextual understanding produces more robust and reliable behavior classification outcomes.

### E. Overall Behavior Classification Performance

This subsection presents the overall performance of the proposed framework for waste disposal behavior classification by integrating YOLOv8 object detection, object tracking, FSM-based temporal reasoning, and Swin Transformer contextual analysis. The evaluation combines training convergence analysis, classification consistency, final prediction performance, and computational efficiency to assess the effectiveness of the proposed hybrid approach in

understanding waste disposal behavior.

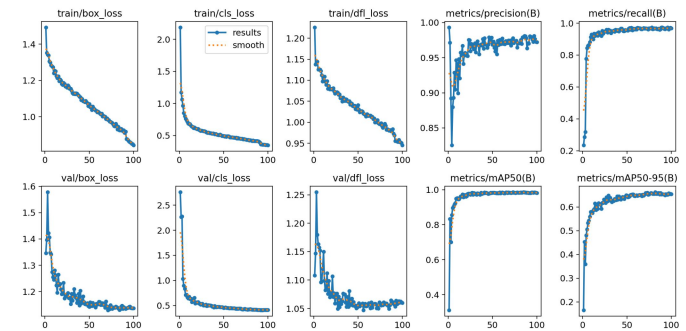


Fig. 10. Model training and validation performance across epochs.

Figure 10 illustrates the training and validation performance of the proposed model across 100 training epochs. Loss convergence, including box loss, classification loss, and distribution focal loss (DFL), shows a steady decrease with both training and validation curves stabilizing at low values, indicating successful learning of spatial and behavioral features without overfitting. Specifically, the training losses decrease from 1.42 to 0.88 (box), 2.10 to 0.70 (classification), and 1.20 to 0.95 (DFL), while the validation losses exhibit similar convergence behavior.

The close agreement between the training and validation curves indicates that the proposed framework generalizes well to unseen data rather than memorizing the training samples. Moreover, the stable convergence of precision, recall, and mAP throughout the training process demonstrates that integrating YOLOv8 with FSM and Swin Transformer does not introduce optimization instability. Instead, the hybrid architecture successfully learns complementary spatial, temporal, and contextual representations, providing a reliable foundation for accurately classifying waste disposal behavior under diverse surveillance conditions.

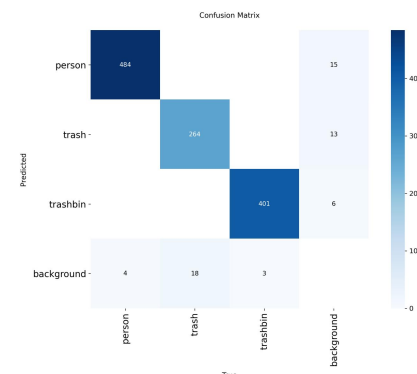


Fig. 11. Confusion matrix for classification accuracy.

The normalized confusion matrix illustrates the predictive accuracy of the proposed framework across different object categories. Correctly classified instances are concentrated along the diagonal, with 484 person instances, 264 trash instances, and 401 trash bin instances correctly recognized.

Misclassifications are limited, including 4 person, 18 trash, and 3 trash bin instances incorrectly predicted as background.

Among all object categories, the trash class exhibits the highest number of misclassifications. This result is expected because waste objects generally occupy a smaller image area, present greater variations in shape and appearance, and are frequently affected by partial occlusion during disposal activities. In contrast, persons and trash bins have relatively stable visual characteristics, resulting in higher classification accuracy. Nevertheless, the low level of inter-class confusion indicates that the integration of spatial detection, temporal tracking, and contextual reasoning effectively reduces classification ambiguity, thereby providing reliable object information for subsequent behavior analysis.

Table 5.  
Results of Model Performance Evaluation

| Object Class       | Precision | Recall | F1-Score | mAP  |
|--------------------|-----------|--------|----------|------|
| Person             | 0.95      | 0.92   | 0.93     | 0.96 |
| Trash              | 0.88      | 0.84   | 0.86     | 0.89 |
| Trash Bin          | 0.93      | 0.90   | 0.91     | 0.94 |
| All Classes (Mean) | 0.92      | 0.88   | 0.90     | 0.93 |

Table 5 summarizes the quantitative performance of the proposed framework. Person detection achieved the highest performance with a Precision of 0.95 and an mAP of 0.96, whereas the trash class obtained comparatively lower scores (Precision = 0.88 and mAP = 0.89).

The lower performance of the trash class can be attributed to the inherent characteristics of waste objects, which often appear smaller, exhibit greater appearance variability, and undergo frequent occlusion during disposal activities. Although these factors increase detection difficulty, the proposed framework maintains a mean Precision of 0.92, Recall of 0.88, F1-score of 0.90, and mAP of 0.93, indicating that the combined use of temporal reasoning and contextual analysis effectively compensates for limitations in object detection. This result demonstrates that reliable behavior classification is achieved not solely through accurate object recognition but also through the integration of temporal interaction modeling and contextual scene understanding. Consequently, the proposed framework is capable of distinguishing proper disposal from littering more consistently than approaches relying only on frame-level object detection.

Besides classification accuracy, computational efficiency was evaluated to determine the applicability of the proposed framework in real-time surveillance environments. Since the framework incorporates object detection, object tracking, FSM-based temporal reasoning, and Swin Transformer contextual analysis, it is important to verify that the additional behavior analysis does not significantly reduce processing speed. Table 6 summarizes the computational performance of the proposed framework.

Table 6.  
Real-Time Performance Evaluation of The Proposed Framework.

|                                 |               |
|---------------------------------|---------------|
| YOLOv8 inference time           | 32.8 ms/frame |
| YOLOv8 inference speed          | 30.5 FPS      |
| Tracking time                   | 3.1 ms/frame  |
| FSM processing time             | 0.6 ms/frame  |
| Swin Transformer inference time | 10.4 ms/frame |
| Total end-to-end latency        | 46.9 ms/frame |
| End-to-end processing speed     | 21.3 FPS      |

Table 6 summarizes the computational performance of the proposed framework for real-time waste disposal behavior classification. The YOLOv8 detector achieved an average inference time of 32.8 ms/frame, corresponding to 30.5 FPS, indicating that the object detection module is capable of real-time processing. After integrating object tracking, FSM-based temporal reasoning, and Swin Transformer contextual analysis, the total end-to-end latency increased to 46.9 ms/frame, resulting in an overall processing speed of 21.3 FPS. The additional latency is primarily introduced by the Swin Transformer (10.4 ms/frame), which performs contextual feature extraction to improve behavior interpretation under complex surveillance conditions. In contrast, the FSM contributes only 0.6 ms/frame, demonstrating that temporal state reasoning imposes negligible computational overhead. Although the proposed framework operates at a lower frame rate than the object detector alone, the reduction in processing speed is compensated by improved contextual reasoning and more reliable behavior classification. Overall, these results demonstrate that the proposed framework achieves an effective balance between computational efficiency and behavior classification accuracy, making it suitable for practical deployment in intelligent smart campus surveillance systems.

#### F. Error Analysis and Limitations

Although the proposed hybrid framework achieves high performance in waste disposal behavior classification, several challenging scenarios may still lead to misclassification. The first failure case occurs during rapid disposal actions, where the waste object is released and disappears within only a few consecutive frames. Under these conditions, the FSM may not accumulate sufficient temporal evidence within the predefined observation window, resulting in a delayed state transition or an incorrect classification of the disposal behavior. A second challenging scenario involves multiple individuals interacting near the same trash bin. When more than one person is in proximity to a single waste object, the tracking module may temporarily assign the object to an incorrect identity, which subsequently affects the FSM state transitions and the final behavior decision.

Additional errors were observed in scenes containing partial occlusion, cluttered backgrounds, or extreme illumination changes. For example, when the released waste object is temporarily occluded by the person or partially hidden by the trash bin, the FSM may continue maintaining the Held state until the object becomes visible again. Similarly, strong shadows or low-light conditions may reduce the quality of the contextual features extracted by the Swin Transformer, leading to uncertainty in distinguishing between proper disposal and

littering when the waste object is located near the boundary of the trash bin.

Despite these challenges, the proposed framework remains robust because temporal reasoning and contextual analysis complement each other. The FSM preserves behavioral consistency through state-transition modeling, whereas the Swin Transformer provides additional contextual information to resolve ambiguous spatial relationships. Consequently, most temporary tracking errors or visual ambiguities can be corrected before the final behavior classification is produced. Nevertheless, the current framework was evaluated using data collected from a single smart campus environment. Future work will investigate adaptive observation windows, more robust multi-object tracking algorithms, and larger multi-campus datasets to further improve generalization and performance under highly dynamic surveillance conditions.

## V. CONCLUSION

This study proposed a hybrid framework for waste disposal behavior classification by integrating YOLOv8 for real-time object detection, object tracking for temporal continuity, a finite state machine (FSM) for interpretable temporal reasoning, and a Swin Transformer for contextual scene understanding. The proposed framework achieved a behavior classification mAP@50 of 0.93, demonstrating that the integration of spatial detection, temporal reasoning, and contextual analysis enables reliable behavior recognition in surveillance environments. The effectiveness of the proposed method lies in the complementary functions of its components: YOLOv8 accurately detects relevant objects, object tracking preserves interaction continuity, the FSM transforms object trajectories into interpretable behavioral states, and the Swin Transformer resolves contextual ambiguities caused by occlusion, complex backgrounds, and uncertain object placement. As a result, the proposed framework overcomes the limitations of conventional object detection approaches that analyze objects independently in each frame, enabling more robust discrimination between proper disposal and littering behaviors.

From a methodological perspective, this study advances behavior-aware surveillance by integrating explicit temporal state-transition modeling with transformer-based contextual reasoning within a unified architecture. Unlike previous approaches that mainly emphasize object recognition or trajectory tracking, the proposed framework combines spatial, temporal, and contextual information to infer behavioral semantics, providing more reliable and interpretable classifications of waste disposal behavior. These capabilities can support smart campus management by automatically monitoring waste disposal practices, identifying littering incidents, evaluating user compliance with waste disposal policies, and providing decision support for campus cleanliness and sustainability programs. Furthermore, the proposed architecture can be extended to smart city applications,

including automated public-space cleanliness monitoring, intelligent environmental surveillance, and urban waste management systems that require continuous behavior-aware monitoring.

Despite these promising results, the proposed framework was evaluated using surveillance data collected from a single campus environment, and the FSM employs a fixed observation window that may be less adaptive to rapid disposal actions or highly dynamic multi-person interactions. Future work will investigate multi-camera fusion to maintain object continuity across different viewpoints, transformer-based temporal modeling as an adaptive alternative to the rule-based FSM, optical-flow-assisted motion analysis for rapid disposal events, and advanced multi-person interaction modeling to improve scalability, robustness, and real-world deployment in large-scale smart campus and smart city surveillance systems.

## REFERENCES

- [1] M. A. Budihardjo, N. G. Humaira, S. A. Putri, B. S. Ramadan, S. Syafrudin, and E. Yohana, "Sustainable solid waste management strategies for higher education institutions: Diponegoro university, indonesia case study," *Sustainability (Switzerland)*, vol. 13, no. 23, Dec. 2021, doi: 10.3390/su132313242.
- [2] E. Longo, F. A. Sahin, A. E. C. Redondi, P. Bolzan, M. Bianchini, and S. Maffei, "A 5g-enabled smart waste management system for university campus," *Sensors*, vol. 21, no. 24, Dec. 2021, doi: 10.3390/s21248278.
- [3] Y. Oba and H. Doi, "Accelerating ecosystem monitoring through computer vision with deep metric learning," Jun. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.ecocom.2025.101124.
- [4] D. Nasien, M. H. Adiya, M. Farkhan, U. S. Rahmadhani, and A. A. Samah, "Automated waste classification using YOLOv11 a deep learning approach for sustainable recycling," *Journal of Applied Business and Technology*, vol. 6, no. 1, pp. 68–74, Jan. 2025, doi: 10.35145/jabt.v6i1.205.
- [5] Z. Wang, Y. Chen, Y. Gu, J. Liu, X. Zhu, and M. He, "The evolution of object detection from CNNs to transformers and multi-modal fusion," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-37052-6.
- [6] S. Uddin, B. Hussain, N. Ahmad, A. Hussain, and S. Fareed, "A hybrid framework for temporal object behavior analysis using LSTM and real-time detection," *International Journal of Ethical AI Application*, no. 4, 2025, doi: https://doi.org/10.64229/91wd9131.
- [7] H. A. Ahmed, S. M. M. Ibrahim, M. S. Kumar, and S. S. Ahmed, "Yolo-Based fast and accurate object detection for real-time applications," *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE*, vol. 2024, no. 23s, pp. 2375–2381, 2024. https://ijisae.org/index.php/IJISAE/article/view/7344
- [8] A. Atalarais, K. S. S. H. Syahputra, A. Idrus, and I. Taufik, "Automatic waste type detection using YOLO for waste management efficiency," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 2, pp. 2808–4519, 2025, doi: 10.59934/jaica.v4i2.770.
- [9] A. M. Rekavandi, S. Rashidi, F. Boussaid, S. Hoefs, E. Akbas, and M. bennamoun, "Transformers in small object detection: A benchmark and Survey of State-of-the-Art," *ACM Journal Computing Surveys*, Sep. 2023, doi: 10.1145/3758090.
- [10] N. Nurmisba, D. Anggreani, M. A. M. Hayat, and A. A. Rahman, "Swin transformer enhanced with OOD detection for robust and reliable diagnosis of ischemic stroke from CT image," *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, vol. 14, no. 3, pp. 502–514, Dec. 2025, doi: 10.23887/janapati.v14i3.102864.
- [11] Lyuyang Gao, "CNN-transformer hybrid models for object detection: A comprehensive review," *Science and Technology of Engineering, Chemistry and Environmental Protection*, no. 4, doi: 10.61173/19bc6r78.

- [12] A. Paul, R. Raj, M. K. Gourisaria, A. V. Jha, and N. Bizon, "YOLO-HARVEST: A hybrid ViT architecture with locality-enhanced attention for automated wildlife species classification," *Ecol. Inform.*, p. 103605, Mar. 2026, doi: 10.1016/j.ecoinf.2026.103605.
- [13] A. Subki, M. Zulpahmi, and B. Imran, "A Hybrid Framework Based on YOLOv8 and Vision Transformer for Multi-Class Detection and Classification of Coffee Fruit Maturity Levels," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 5, pp. 2548–6861, 2025, doi: 10.30871/jaic.v9i5.10590.
- [14] F. Fathurahman, D. Indra, T. Hasanuddin, H. Darwis, and T. Kazuaki, "Smart waste bin prototype for university waste management," *Indonesian Journal of Data and Science*, vol. 6, no. 3, pp. 378–392, Dec. 2025, doi: 10.56705/ijodas.v6i3.324.
- [15] M. W. Rahman, R. Islam, A. Hasan, N. I. Bithi, M. M. Hasan, and M. M. Rahman, "Intelligent waste management system using deep learning with IoT," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 5, pp. 2072–2087, May 2022, doi: 10.1016/j.jksuci.2020.08.016.
- [16] F. R. Sayem *et al.*, "Enhancing waste sorting and recycling efficiency: robust deep learning-based approach for classification and detection," *Neural Comput. Appl.*, vol. 37, no. 6, pp. 4567–4583, Feb. 2025, doi: 10.1007/s00521-024-10855-2.
- [17] Z. Kang, J. Yang, G. Li, and Z. Zhang, "An Automatic Garbage Classification System Based on Deep Learning," *IEEE Access*, vol. 8, pp. 140019–140029, 2020, doi: 10.1109/ACCESS.2020.3010496.
- [18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," doi: 10.48550/arXiv.1506.02640.
- [19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOV4: Optimal speed and accuracy of object detection," *arXiv (Cornell University)*, Apr. 2020, doi: 10.48550/arxiv.2004.10934.
- [20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-End Object Detection with Transformers*. Springer International Publishing, 2020. doi: 10.48550/arXiv.2005.12872.
- [21] G. Maroun, S. E. Bekhouche, J. Charafeddine, and F. Dornaika, "Integrating ConvNeXt and vision transformers for enhancing facial age estimation," *Computer Vision and Image Understanding*, vol. 262, Dec. 2025, doi: 10.1016/j.cviu.2025.104542.
- [22] J. Yang, H. Wan, and Z. Shang, "Enhanced hybrid CNN and transformer network for remote sensing image change detection," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-94544-7.
- [23] Y. Ye, Q. Sun, K. Cheng, X. Shen, and D. Wang, "A lightweight mechanism for vision-transformer-based object detection," *Complex and Intelligent Systems*, vol. 11, no. 7, Jul. 2025, doi: 10.1007/s40747-025-01904-x.
- [24] S. J. Hadi, I. Ahmed, A. Iqbal, and A. S. Alzahrani, "CATR: CNN augmented transformer for object detection in remote sensing imagery," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-27872-3.
- [25] A. De, A. P. Amit Nigam, and S. Raman Chatterjee, "Design and Simulation of an Optimized Traffic Controller Using Moore FSM," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 12, pp. 1047–1060, Dec. 2024, doi: 10.22214/ijraset.2024.65958.
- [26] M. B. R. Alvian, S. Bukhori, and M. A. Furqon, "Implementation of finite state machine to determine the behaviour of non-playable character in leadership simulation game," *Journal of Games Game Art and Gamification*, vol. 9, no. 1, pp. 11–21, Jun. 2024, doi: 10.21512/jggag.v9i1.10894.
- [27] M. Talib, A. H. Y. Al-Noori, and J. Suad, "YOLOv8-CAB: Improved YOLOv8 for real-time object detection," *Karbala International Journal of Modern Science*, vol. 10, no. 1, pp. 56–68, 2024, doi: 10.33640/2405-609X.3339.
- [28] L. Rodriguez-Benitez, C. Solana-Cipres, J. Moreno-Garcia, and L. Jimenez-Linares, "Approximate reasoning and finite state machines to the detection of actions in video sequences," *International Journal of Approximate Reasoning*, vol. 52, no. 4, pp. 526–540, Dec. 2010, doi: 10.1016/j.ijar.2010.12.003.
- [29] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv (Cornell University)*, Mar. 2021, doi: 10.48550/arxiv.2103.14030.
- [30] I. A. Kandhro *et al.*, "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification," *Heliyon*, vol. 10, no. 10, May 2024, doi: 10.1016/j.heliyon.2024.e31488.