Vision-Based Surveillance for Malpractice Detection: An Analysis of Pose Estimation and Object Detection

Aashish Tom Raju
Deparment Of CSE
Amal Jyothi College of Engineering
Kottayam, India
aashishtomraju2026@cs.ajce.in

Aneesh Varghese John
Deparment Of CSE
Amal Jyothi College of Engineering
Kottayam, India
aneeshvarghesejohn2026@cs.ajce.in

Ashish Shabu
Deparment Of CSE
Amal Jyothi College of Engineering
Kottayam, India
ashishshabu2026@cs.ajce.in

Bibin Babu
Deparment Of CSE
Amal Jyothi College of Engineering
Kottayam, India
bibinbabu2026@cs.ajce.in

Ms. Anishamol Abraham
Deparment Of CSE
Amal Jyothi College of Engineering
Kottayam, India
anishamolabraham@amaljyothi.ac.in

***Abstract*—This paper introduces a smart, real-time surveillance system for malpractice detection by combining advanced object detection and human pose estimation. The object detection part uses established techniques along with Convolutional Neural Networks (CNNs) to effectively identify and track objects in video streams. This method has proven highly effective, demonstrating top-tier accuracy (MOTA) and precision (MOTP) on benchmark datasets like MOT16 and MOT17.**

To analyze human behavior, the system employs an enhanced YOLOv8 model for pose estimation, which improves both speed and accuracy. This model features two key upgrades: a SimDLKA attention mechanism to better focus on medium-to-large targets and a new DCIOU loss function that makes training more stable and efficient. These improvements result in a 2.7% boost in mAP (mean Average Precision) and faster frame rates on standard datasets like COCO and MPII. This combined platform provides a robust solution for monitoring in complex environments, suitable for applications from security and traffic control to advanced human motion analysis.

Keywords*— *Malpractice Detection, Human Pose Estimation, Object Detection, YOLOv8, Intelligent Surveillance Systems, Real-time Monitoring, Academic Integrity, Computer Vision, Deep Learning, YOLOv8-Pose, MediaPipe, Gesture Recognition, FaceNet, Facial Emotion Recognition (FER), Multi-modal AI, Automated Invigilation, Behavioral Analysis, Decision Engine.

I. INTRODUCTION

Maintaining academic integrity is a cornerstone of quality education, yet ensuring it in traditional examination settings presents significant challenges. Manual invigilation in crowded exam halls is resource-intensive and prone to human error, making it difficult to monitor every student effectively. While existing automated solutions have been proposed, they often lack the capability for real-time feedback or fail to detect the wide range of malpractice methods employed by students. This creates a critical need for an intelligent system that can offer comprehensive, real-time monitoring to ensure a fair and secure examination process.

Common forms of academic malpractice range from the use of prohibited items, such as mobile phones and unauthorized notes, to non-verbal communication through suspicious head movements and hand gestures. An effective automated system must therefore be capable of not just identifying objects but also interpreting human behavior. This requires a sophisticated, multi-faceted approach that goes beyond simple video monitoring to address the diverse and subtle ways in which cheating can occur.

To address these limitations, we propose InvigiLens, an intelligent, multi-modal AI system designed to automate the detection and monitoring of exam malpractice in offline environments. The system integrates state-of-the-art object detection and pose estimation to identify prohibited items, suspicious gestures, and non-verbal communication between students. The main contribution of this work is the development of a multi-modal system specifically aimed at practical offline classroom scenarios. It features a real-time alert and screenshot system that instantly flags suspicious activities to provide evidence, ultimately creating an invigilator-free detection framework that reduces the need for constant human supervision. Through this system, we demonstrate a robust and scalable solution that significantly enhances the fairness, transparency, and security of the examination process.

II. LITERATURE REVIEW

A. Real-time object detection, tracking, and monitoring framework for security surveillance systems

The growing global need for robust security has intensified the demand for automated surveillance systems capable of real-time object detection, tracking, and monitoring. In response, the research by Abba *et al.* presents a comprehensive software framework designed specifically for security surveillance applications. The study's primary objective is to provide a flexible and dynamically comprehensible system that can be deployed in real-world situations to aid security personnel. The proposed solution is a hybrid model that integrates established computer vision algorithms—including background subtraction, approximate median filtering, and component labeling—with a Convolutional Neural Network (CNN) for high-level validation and control.

The framework's architecture is a multi-stage pipeline designed for both efficiency and accuracy. The initial object detection phase uses a Background Subtraction Algorithm (BSA), a well-established method for identifying moving objects in video sequences by comparing each frame to a static reference background model. To address the image noise common in surveillance footage, this stage incorporates an Approximate Median Filter (AMF), a non-linear algorithm that is highly effective at reducing impulse noise (like salt-and-pepper) while preserving critical edge details of the detected objects. Once foreground objects are successfully segmented, the system employs a Connected-Component Labeling (C-CL) algorithm for the tracking phase. This technique assigns a unique label to all pixels belonging to a single object, which is crucial for differentiating and tracking multiple objects as they move across consecutive frames. A Convolutional Neural Network (CNN) serves as the high-level intelligence layer, orchestrating the entire workflow through validation, classification, and recognition tasks. The CNN validates the outputs from the detection and tracking modules and is utilized for more complex operations such as face recognition to identify specific individuals. This hybrid design leverages the computational speed of classical algorithms for initial processing while harnessing the powerful pattern recognition capabilities of deep learning for final analysis. The framework's performance was rigorously evaluated on the Multiple Object Tracking (MOT) Challenge benchmarks, specifically the MOT15, MOT16, and MOT17 datasets. On the MOT15 dataset, the system achieved a Multiple Object Tracking Accuracy (MOTA) of 53.9%, outperforming several state-of-the-art methods. Similarly, it recorded a MOTA of 53.7% on the MOT17 dataset, again demonstrating superior performance. These consistently strong results across challenging benchmarks confirm that the framework is a robust and effective solution suitable for deployment in real-world security and surveillance systems.

B. A Human Pose Estimation Network Based on the YOLOv8 Framework

Deep learning-based human pose estimation is a core task in computer vision, but traditional methods for multi-person scenarios often struggle with challenges like partial occlusions and overlaps between human bodies. To address these issues, the research by Cai et al. proposes a new human pose estimation network named EE-YOLOv8, which is built upon the YOLOv8 framework. The proposed model integrates three key innovations to improve performance: an Efficient Multi-scale Receptive Field (EMRF) module, an Expanded Feature Pyramid Network (EFPN), and the Wise-IoU loss function. The primary goal is to achieve a superior balance between sub-pixel localization accuracy and real-time processing efficiency, especially in complex environments.

The architectural and functional enhancements of EE-YOLOv8 are threefold. First, the EMRF module is employed to improve the model's feature representation capability by replacing the original C2f module in the backbone network. This module incorporates an Efficient Multi-scale Attention (EMA) mechanism, which enhances the network's focus on human-related features and improves its ability to understand objects at different scales. Second, an Expanded Feature Pyramid Network (EFPN) is introduced to replace the original Path Aggregation Network (PAN) in the model's neck. The EFPN optimizes information exchange between different feature levels and enhances multi-scale data integration by incorporating more top-down and bottom-up paths, as well as the detail-rich P2 feature layer. Finally, the model replaces the traditional Intersection over Union (IoU) with the Wise-IoU loss function to improve detection accuracy. Wise-IoU uses a dynamic non-monotonic focusing mechanism to better handle low-quality examples and complex scenes with overlapping objects, which also helps to accelerate the model's convergence speed.

The proposed EE-YOLOv8 model was evaluated on the large-scale MS COCO 2017 dataset, which contains over 200,000 images and 250,000 human instances. Compared to the baseline YOLOv8-Pose, EE-YOLOv8 demonstrated significant improvements, achieving an Average Precision (AP) of 89.0 % at an IoU threshold of 0.5 (a 3.3% increase) and an AP of 65.6% over the IoU range of 0.5-0.95 (a 5.8% increase). The authors note that EE-YOLOv8 achieves this higher accuracy while having the lowest parameter count among the analyzed state-of-the-art algorithms. Ablation experiments confirmed that all three integrated components contributed to the performance gains, with the EFPN being the most impactful innovation. Visual analysis further showed that while performance in single-person scenes was similar, EE-YOLOv8 had significantly fewer missed or false detections in multi-person scenes compared to the baseline model.

III. METHODOLOGY

A. System Architecture

This section details the design and implementation of our proposed system, InvigiLens, a real-time, multi-modal frame-

work for automated exam malpractice detection. The system is built upon the robust YOLOv8 framework, which is leveraged for both its high-speed object detection and its sophisticated pose estimation capabilities. The following subsections will first present the overall system architecture, then provide a detailed explanation of the individual modules for pose estimation and object detection,and finally, describe the rule-based logic that integrates the outputs from these modules to accurately identify and flag potential malpractice activities.

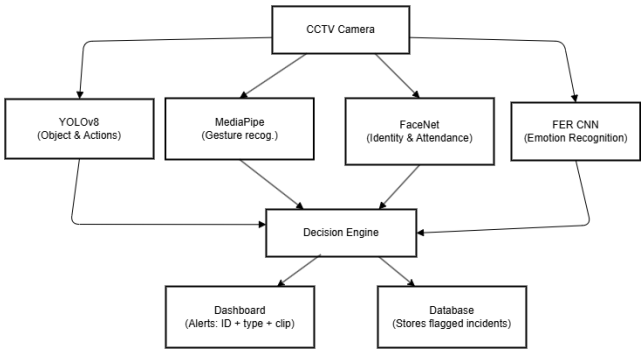

Fig. 1. System Architecture

Following pre-processing, each frame is broadcast simultaneously to four parallel deep learning modules, each specialized for a distinct analytical task. This parallel processing ensures that multiple facets of student behavior are analyzed efficiently without creating a computational bottleneck.

The system utilizes a YOLOv8 module for detecting prohibited objects like mobile phones or notes and classifying high-level student actions. Alongside this, a MediaPipe module is dedicated to fine-grained gesture recognition, identifying specific hand or body movements that could indicate cheating. A FaceNet module handles identity verification and attendance; this crucial component ensures that any flagged incident is correctly associated with a specific student. Finally, a Facial Emotion Recognition (FER) CNN module analyzes facial expressions to infer emotional states, providing supplementary data that may correlate with malpractice.

The outputs from all four analytical modules are then aggregated and fed into a central Decision Engine. This engine serves as the core logic of the system, applying a set of predefined rules to the combined data streams: objects, gestures, identity, and emotions to intelligently determine if a malpractice event has occurred. If the Decision Engine flags an incident, it triggers two actions in the output stage: an immediate alert is generated and sent to a real-time Dashboard, containing the student's ID, the type of malpractice, and a short video clip for invigilator verification. Concurrently, the incident details are logged and stored in a Database for record-keeping and future review.

B. Core Technology: YOLOv8 Framework

The core of our system is built upon the **YOLOv8** (You Only Look Once, version 8) framework, a state-of-the-art, real-time object detection algorithm. Its key advantage lies in its end-to-end design, which processes an entire image and produces detections in a single forward propagation through the network. This approach simplifies the entire processing pipeline and significantly improves overall efficiency, making the framework an ideal choice for the real-time requirements of a live exam monitoring application like InvigiLens.

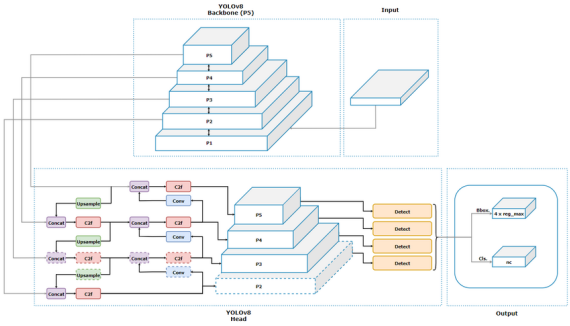

Fig. 2. YOLOv8 Architecture

The YOLOv8 architecture is a highly efficient and scalable model designed for real-time object detection. It is logically divided into three main parts: the Backbone, the Neck, and the Head. Each component is optimized for its specific task, working together to transform an input image into a set of bounding boxes and class predictions. The overall structure is designed to be anchor-free, which simplifies the training process and improves detection speed.

1) Backbone: The Backbone is primarily responsible for extracting feature hierarchies from the input image. It processes the initial 640x640 pixel image through a series of convolutional layers. The core building block of the YOLOv8 backbone is the C2f module, which is an evolution of the C3 module found in YOLOv5. The C2f module promotes richer gradient flow by splitting the feature channels and processing them through a series of lightweight Bottleneck blocks before concatenating them. This design allows for more effective feature extraction. The backbone progressively downsamples the spatial dimensions of the feature maps while increasing their channel depth, creating feature maps at various scales (P1 through P5). The final stage of the backbone employs a Spatial Pyramid Pooling - Fast (SPPF) module. The SPPF module uses several sequential 5x5 max-pooling layers to create a fixed-size output, effectively increasing the network's receptive field and allowing it to capture contextual information at different scales without a significant increase in computation.

2) Neck: The Neck of the network is designed to fuse the features extracted by the backbone. Its primary purpose is to combine feature maps from different scales to create a rich, multi-scale feature representation that is robust for detecting objects of various sizes. YOLOv8 employs a Path Aggregation Network (PAN) integrated with a Feature Pyramid Network

(FPN).

Top-Down Pathway (FPN): The neck first takes the high-level feature maps from the deeper parts of the backbone (e.g., P3, P4, P5) and upsamples them. It then merges these semantically rich features with the spatially finer features from earlier in the backbone.

Bottom-Up Pathway (PAN): After the top-down fusion, a second, bottom-up pathway is introduced. This path propagates the strong positional information from the lower-level features back up to the higher-level feature maps, further enhancing the localization accuracy for all scales.

3) Head: The Head is the final component of the network, responsible for generating the actual object detections from the fused features provided by the neck. The YOLOv8 Head is both anchor-free and decoupled.

Anchor-Free: Unlike previous models that relied on predefined anchor boxes, the YOLOv8 head directly predicts the center of an object along with its height and width. This reduces the number of predictions and simplifies the post-processing steps like Non-Max Suppression (NMS).

Decoupled: The head uses separate convolutional layers for the classification and regression tasks. One branch predicts the class probabilities for the detected object, while another branch regresses the bounding box coordinates. This separation resolves the conflict between the two tasks and has been shown to improve overall model accuracy. The head outputs predictions at three different scales (corresponding to feature maps of sizes 80x80, 40x40, and 20x20), allowing it to effectively detect small, medium, and large objects within the same image. The final loss is calculated using a combination of a classification loss (Cls. Loss) and a bounding box regression loss (Bbox. Loss), which often includes Distribution Focal Loss (DFL) for more precise localization.

C. Pose Estimation Module

To monitor and analyze student posture and actions, our system employs the YOLOv8-Pose variant. This model extends the base YOLOv8 architecture by incorporating an additional output head specifically designed for human keypoint detection. This modification allows the network to perform two tasks simultaneously in a single forward pass: it first detects each person within the frame and then localizes their anatomical keypoints, such as the head, shoulders, elbows, and wrists. This integrated, end-to-end approach is highly efficient and provides the detailed skeletal data required for interpreting complex human behaviors in real-time.

Within the InvigiLens framework, the Pose Estimation Module is trained to recognize several specific behaviors that are indicative of potential malpractice. The system analyzes the stream of keypoint data to identify and flag the following patterns: The module monitors for **Abnormal Head Turning** by continuously tracking the orientation of each student's head; if a student's head is turned to the side to look at a neighbor's desk for a sustained period, the system flags it as a suspicious event.

The model is also trained to detect **Suspicious Gestures**, which are specific hand and arm movements not typical in an exam setting, such as gestures used for non-verbal communication or movements associated with accessing concealed notes.

Lastly, to differentiate between malpractice and a legitimate request for help, the system is trained to recognize the distinct action of **Hand Raising**, which allows it to notify the invigilator of a student needing assistance without generating a false malpractice alert.

D. Object Detection Module

In parallel with the pose estimation module, the system employs an Object Detection Module to identify prohibited items within the examination environment. This module is implemented using the standard YOLOv8 object detection framework. We selected YOLOv8 for this task due to its proven high accuracy and exceptional inference speed, which are critical for processing a live video stream without delay. The module analyzes each frame to locate and place a bounding box around any object belonging to a predefined set of target classes associated with academic dishonesty.

The model was fine-tuned on a custom-curated dataset to specialize in recognizing items frequently used for exam malpractice, and the specific target classes are Mobile Phones, Smartwatches, Paper Notes / Chits, and Other portable electronic devices, the detection of one or more of which serves as a key input for the system's Decision Engine.

IV. EXPERIMENTS AND RESULTS

This section outlines the experimental setup and presents a preliminary performance analysis of the proposed InvigiLens system. The primary objective of this evaluation is to validate the feasibility of our multi-modal approach and to establish a performance baseline for the system's accuracy and real-time processing capabilities. The system was tested using a curated dataset containing scenarios relevant to an examination environment. In the following subsections, we describe the dataset composition, detail the experimental setup and evaluation metrics, and present the initial results of our framework.

A. Dataset

For this study, a specialized dataset was curated to train and evaluate our 3791system. To ensure a diverse range of scenarios, lighting conditions, and subjects, data was aggregated from several relevant, publicly available datasets on the Kaggle platform.

The primary Kaggle datasets used as sources were:

- *Aggregate_Dataset_Exam_cheatDetection* by Bhishm Bhandari
- *exam cheating detection* by Shaon
- *Classroom Student Behaviors* by phamluhuynhmai
- *ExamCheating_Dataset* by Ardutra Agi Ginting

From these sources, a manual curation process was performed to select, clean, and re-annotate images and video

frames specifically relevant to the malpractice behaviors defined in our methodology. The final curated dataset consists of 3,000 images, which were categorized into four main classes to train our multi-modal detection logic. The classes include: Phone Usage, Looking Around, Leaning to copy, **Sharing Notes** and a **Normal** behavior class.

This final dataset was then randomly partitioned into a 70% training set, a 15% validation set, and a 15% testing set. This split ensures that the model's performance is evaluated on data it has not seen during the training or tuning phases.

B. Experimental Setup

All experiments were conducted on a laptop equipped with an Intel Core i7-12700H CPU, 16 GB of DDR4 RAM, and an NVIDIA GeForce RTX 3060 Laptop GPU with 6 GB of VRAM. The software environment consisted of the Windows 11 operating system, with the models being implemented and trained using the PyTorch deep learning framework (version 1.13). GPU acceleration was enabled using CUDA 11.7, and key libraries such as OpenCV were used for image processing. For training the model, we used a batch size of 16 and trained for 100 epochs. The initial learning rate was set to 0.01, and we used the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 to update the model's parameters.

C. Evaluation Metrics

To assess the performance of our proposed model, we utilized a set of standard metrics from the field of object detection. The model's accuracy was primarily evaluated using Precision, Recall, and mean Average Precision (mAP). Precision measures the accuracy of the positive predictions, indicating how many of the detected malpractice incidents were correct. Recall measures the model's ability to identify all actual malpractice instances within the dataset. The primary metric for overall performance is the mAP@0.5, which calculates the average precision at an Intersection over Union (IoU) threshold of 50%. Since the current implementation focuses on detecting malpractice from static images, we measure computational performance by the average inference time per image. This metric serves as a key indicator of the system's efficiency and its feasibility for deployment in a future real-time video processing pipeline, where low inference time per image directly translates to a high number of frames per second (FPS).

D. Performance Analysis

1) Quantitative Results: To validate the performance of our framework, the trained model was evaluated on the held-out test set to assess its ability to correctly classify the type of malpractice occurring in a given image. The system achieved a top-1 accuracy of 87.2%, which indicates that the model's single highest-confidence prediction correctly identified the specific malpractice type in over 87% of cases. In terms of computational performance, the model demonstrated high efficiency with an average inference time of 46.7 milliseconds per image. This rapid processing speed confirms the suitability of the YOLOv8 framework for our intended real-time application, as it corresponds to a potential processing speed of over 21 frames per second.

E. Qualitative Analysis

To provide a visual demonstration of the InvigiLens system's capabilities, this section presents several representative screenshots of the framework detecting various forms of malpractice, as well as normal behavior, in simulated examination environments. These qualitative examples illustrate the model's ability to accurately classify student actions and serve to highlight the practical application and effectiveness of our proposed system.

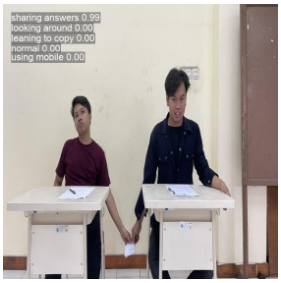

Fig. 3. System detection of a "sharing answers" event.

In Fig. 3, the system successfully identifies "sharing answers" with a high confidence score of 0.99. This demonstrates the model's ability to detect subtle interactions between students that indicate collaboration or the exchange of information, likely through the passing of a note.

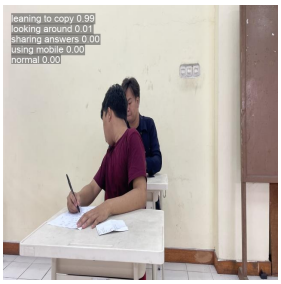

Fig. 4. System detection of a "leaning to copy" event.

In Fig. 4, the system accurately flags "leaning to copy" with a confidence of 0.99. This detection is crucial for identifying instances where a student attempts to view another's exam paper, highlighting the model's sensitivity to specific body postures and proximity.

Fig. 5 illustrates the detection of "looking around" behavior with a confidence score of 0.74. While not as direct an act of cheating as copying, this behavior is often a precursor or an indicator of potential malpractice, showcasing the system's ability to identify suspicious vigilance.

Fig. 6 showcases the detection of "using mobile" with a confidence of 1.00. This is a critical capability for any proctoring system, as mobile phones are a primary tool for unauthorized information access. The model's high confidence

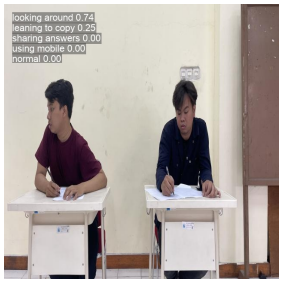

Fig. 5. System detection of a ”looking around” event.

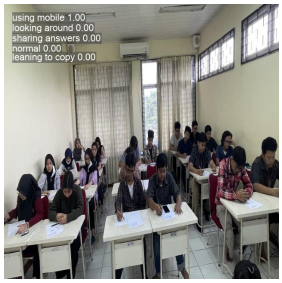

Fig. 6. System detection of a ”using mobile” event.

in this detection underscores its effectiveness in identifying prohibited electronic devices.

Fig. 7. System correctly identifying ”normal” student behavior.

Finally, Fig. 7 presents a scenario of normal student behavior, which the system correctly classifies as ”normal” with a perfect confidence score of 1.00. This example is vital as it demonstrates the model’s robustness in distinguishing legitimate exam conduct from suspicious activities, thereby minimizing false positives.

V. CONCLUSION

This paper introduces an intelligent, multi-modal surveillance system designed to address the significant challenges of manual exam invigilation, which is often resource-intensive and prone to human error. The proposed solution aims to automate the detection of academic malpractice in real-time by integrating state-of-the-art object detection and pose estimation techniques. The system’s architecture is centered on the YOLOv8 framework and operates by feeding CCTV footage into four parallel modules: a YOLOv8 module for detecting prohibited objects (like mobile phones and notes) , a MediaPipe module for gesture recognition , a FaceNet module for student identification , and an FER CNN for emotion recognition. The outputs from these modules are aggregated by a central Decision Engine that applies predefined rules to identify malpractice, such as abnormal head turning or suspicious gestures, and then sends alerts to a dashboard.

To validate this framework, we conducted experiments using a specialized dataset of 3,000 images curated from several public sources and categorized into classes like ”Phone Usage,” ”Leaning to copy,” and ”Normal”. The model achieved a top-1 classification accuracy of 87.2% and a fast average inference time of 46.7 milliseconds per image, demonstrating its capability for real-time processing at over 21 frames per second. Qualitative results further confirmed the system’s effectiveness, showing high-confidence detections for various malpractice activities while also correctly identifying normal student behavior, thereby indicating a low rate of false positives. The paper successfully presents a robust proof-of-concept for an automated invigilation system that can enhance academic integrity.

REFERENCES

[1] Algabri, R., Shin, H., Abdu, A., Bae, J.-H., Lee, S. (2025). WQuatNet: Wide range quaternion-based head pose estimation. Journal of King Saud University – Computer and Information Sciences

[2] Algabri, R., Abdu, A., Lee, S. (2024). Deep learning and machine learning techniques for head pose estimation: a survey. Artificial Intelligence Review,

[3] Rieger, I., Hauenstein, T., Hettenkofer, S., Garbas, J.-U. (2019). Towards Real-Time Head Pose Estimation: Exploring Parameter-Reduced Residual Networks on In-the-wild Datasets.

[4] Wang, J., Zhang, J., Luo, C.,Chen, F. (2017). Joint head pose and facial landmark regression from depth images. Computational Visual Media

[5] Cheng, G., Chao, P., Yang, J., Ding, H. (2024). SGST-YOLOv8: An Improved Lightweight YOLOv8 for Real-Time Target Detection for Campus Surveillance.

[6] Lei, X., Wu, S., Wu, W., Jiang, Z. (2025). MambaNeXt-YOLO: A Hybrid State Space Model for Real-time Object Detection.

[7] Nimma, D., Al-Omari, O., Pradhan, R., Zoirov, U., Krishna, R. V. V., El-Ebiary, T. Y. A. B., Rao, V. S. (2025). Object detection in real-time video surveillance using attention based transformer-YOLOv8 model.

[8] Abba, S., Bizi, A. M., Lee, J.-A., Bakouri, S., Crespo, M. L. (2024). Real-time object detection, tracking, and monitoring framework for security surveillance systems.

[9] Do, V.-D., Le, V.-H., Do, H.-S., Phan, V.-N., Te, T.-H. (2024). TQU-HG dataset and comparative study for hand gesture recognition of RGB-based images using deep learning.

[10] Kapitanov, A., Makhlyarchuk, A., Kvanchiani, K. (2022). HaGRID - HAnd Gesture Recognition Image Dataset.

[11] Mujahid, A., Awan, M. J., Yasin, A., Mohammed, M. A., Damaševičius, R., Maskeliūnas, R., Abdulkareem, K. H. (2021). Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model.

[12] Srivastava, V., Shukla, S., Shyam, R. (2023). Utilizing ML for Hand Gesture Recognition. Journal of Image Processing Pattern Recognition Progress.

[13] Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., and Lu, C. (2022). AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time.

[14] **Figure 2**: Jegham, Nidhal Koh, Chan Young Abdelatti, Marwan Hendawi, Abdeltawab. (2024). Evaluating the Evolution of YOLO (You Only Look Once) Models