SURVEY ON RETAIL INSIGHT GENERATOR: VISION AI FOR CUSTOMER ANALYTICS AND HEATMAPS

Ahmed Razal, Amalkrishna KB, Edwin Davis P, Hariprasad PS.

Ahmed Razal Student, Computer Science and Engineering, Holy Grace Academy of Engineering, Kerala, India

Amalkrishna KB Student, Computer Science and Engineering, Holy Grace Academy of Engineering, Kerala, India

Edwin Davis P Student, Computer Science and Engineering, Holy Grace Academy of Engineering, Kerala, India

Hariprasad PS Student, Computer Science and Engineering, Holy Grace Academy of Engineering, Kerala, India

ABSTRACT

The rapid evolution of retail environments has intensified the need for intelligent systems capable of understanding customer behavior and optimizing in-store operations. Traditional retail analytics methods often lack real-time insights and detailed behavioral tracking, limiting their effectiveness in data-driven decision-making. This study presents a Vision AI-based Retail Insight Generator that leverages advanced computer vision and deep learning techniques—specifically Convolutional Neural Networks (CNNs) and real-time object detection models such as YOLO—to analyze customer activity from video streams. The proposed system is designed to detect, track, and analyze customers within retail spaces, extracting meaningful insights such as footfall distribution, dwell time, movement patterns, and demographic attributes including age and gender. The architecture follows a structured pipeline consisting of video preprocessing, human detection and tracking, feature extraction, and analytics generation. Additionally, the system incorporates heatmap visualization to identify high-engagement zones and an AI-driven strategy engine that provides actionable business recommendations based on observed patterns. To enhance decision-making capabilities, the system includes modules for real-time monitoring, historical trend analysis, and hourly engagement evaluation, enabling retailers to understand customer behavior across both temporal and spatial dimensions. The integration of analytics with an interactive dashboard ensures intuitive visualization and efficient interpretation of complex data. By transforming raw video data into actionable intelligence, the proposed system significantly reduces manual effort and enhances operational efficiency. It is particularly beneficial for retail environments with limited access to advanced analytical tools, providing a scalable and cost-effective solution. The results demonstrate that AI-assisted retail analytics systems can play a crucial role in improving customer experience, optimizing store layouts, and driving strategic business growth. This study highlights the potential of Vision AI as a transformative technology in modern retail analytics and intelligent decision-support systems.

Keyword: - *Retail Analytics, Computer Vision, YOLO, Deep Learning, Object Detection, Customer Behavior, Heatmap, Real-Time Analytics .*

1. INTRODUCTION

The retail industry is undergoing rapid transformation driven by increasing competition, evolving customer expectations, and the growing demand for data-driven decision-making. Understanding customer behavior within physical retail environments remains a significant challenge, as traditional methods such as manual observation and sales data analysis provide limited and delayed insights. Unlike online platforms, which offer detailed user analytics, physical stores often lack real-time visibility into customer interactions, movement patterns, and engagement levels.Recent advancements in artificial intelligence (AI) and deep learning have opened new avenues for intelligent retail analytics through computer vision-based systems. Vision AI enables the analysis of video data captured from in-store surveillance cameras, allowing retailers to extract meaningful insights about customer behavior. This study proposes a Vision AI-based Retail Insight Generator designed to analyze and interpret customer activities using advanced object detection and tracking models, such as Convolutional Neural Networks (CNNs) and real-time detection frameworks like YOLO.The proposed system automates the end-to-end analytics pipeline, including video preprocessing, human detection, multi-object tracking, demographic analysis, and behavioral insight generation. It identifies customers within video streams, tracks their movement across different store zones, and computes key metrics such as footfall, dwell time, and zone-wise engagement. Additionally, the system generates heatmaps to visually represent high-traffic areas, enabling retailers to optimize store layouts and product placement strategies.To further enhance decision-making capabilities, the system incorporates advanced analytical modules such as demographic breakdown analysis, hourly engagement trends, and historical data comparison. An AI-driven strategy engine is also integrated to generate actionable business recommendations based on observed patterns, supporting improved operational planning and customer experience enhancement.By reducing manual effort, minimizing observational bias, and enabling real-time analytics, the proposed system aims to transform traditional retail environments into intelligent, data-driven ecosystems. Its ability to provide scalable, automated, and accurate insights demonstrates its potential for seamless integration into modern retail workflows. Ultimately, this work contributes toward bridging the gap between physical retail and digital analytics, enabling smarter decision-making and improved business outcomes.

2. MILESTONES

The paper *"When AI Meets Store Layout Design: A Review (2022)"* presents a comprehensive study on the application of artificial intelligence in optimizing retail store layouts to enhance customer experience and maximize sales performance. Traditional store layout design relies heavily on manual observation, heuristic methods, and historical sales data, which often fail to capture real-time customer behavior and dynamic interaction patterns within retail spaces. This study highlights how AI-driven approaches can address these limitations by leveraging computer vision and data analytics. The authors explore the integration of CCTV-based surveillance systems with deep learning models to analyze customer movement patterns, dwell time, and interaction zones. By utilizing techniques such as object detection, tracking, and behavioral analysis, the system can identify high-traffic areas and predict customer flow within a store. These insights are then used to recommend optimal product placement, aisle design, and store navigation strategies. The paper also discusses the use of reinforcement learning and predictive analytics to simulate different layout configurations and evaluate their impact on customer engagement and sales. A major strength of this study lies in its extensive review of existing store layout strategies and its practical emphasis on utilizing already deployed surveillance infrastructure, making it cost-effective for retailers. However, the work remains largely theoretical, as it does not include real-world implementation or experimental validation to support its claims. Additionally, the absence of detailed cost–benefit analysis and limited discussion on real-time deployment challenges, such as latency and scalability, restrict its applicability in dynamic retail environments. Despite these limitations, the paper provides valuable insights into the potential of AI-driven layout optimization in modern retail systems.

The research paper *"Deep Learning Based Approach to Detect Customer Age, Gender and Expression in Surveillance Video (2020)"* focuses on extracting demographic and emotional insights from retail surveillance footage to support personalized marketing and customer behavior analysis. The study emphasizes the importance of understanding customer profiles, as demographic attributes such as age and gender, along with emotional states, play a crucial role in influencing purchasing decisions and store engagement. The proposed system utilizes advanced deep learning architectures, including Wide Residual Networks (WideResNet) and Xception-based Convolutional Neural Networks, for robust facial analysis. The methodology involves multiple stages, including face detection, alignment, feature extraction, and classification. The system is designed to operate effectively even on low-resolution surveillance

footage, which is a common challenge in real-world retail environments. In addition to demographic classification, the model incorporates emotion recognition capabilities, enabling retailers to assess customer satisfaction and engagement levels in real time. One of the key strengths of this study is its ability to combine demographic and sentiment analysis into a unified framework, providing a more comprehensive understanding of customer behavior. This integration allows businesses to design targeted marketing campaigns, optimize product placement, and enhance customer experience. However, the system faces several practical challenges. Its performance significantly degrades under conditions such as poor lighting, occlusions, and unfavorable camera angles. Furthermore, the model requires a large volume of labeled training data, making its implementation resource-intensive and potentially costly. Despite these limitations, the study demonstrates the effectiveness of deep learning in extracting meaningful insights from surveillance data and highlights its potential applications in intelligent retail systems.

The paper *"FMViT: A Multiple-Frequency Mixing Vision Transformer (2023)"* introduces an advanced hybrid architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to improve both accuracy and computational efficiency in visual recognition tasks. Traditional CNNs are highly effective in capturing local spatial features, while Vision Transformers excel in modeling global dependencies. FMViT aims to integrate these complementary strengths into a unified framework. The proposed model incorporates innovative components such as frequency-mixing modules and feature fusion blocks. These mechanisms enable the network to process visual information across multiple frequency domains, effectively capturing both fine-grained details and global contextual information. This design significantly enhances the model's ability to handle complex visual tasks such as object detection, tracking, and scene understanding, making it particularly suitable for applications like retail analytics and surveillance systems. One of the major advantages of FMViT is its ability to achieve high accuracy while maintaining efficient computational performance. The model is designed to reduce redundancy in feature extraction and improve latency, making it viable for deployment in resource-constrained environments. Additionally, its hybrid architecture allows it to outperform traditional CNN-based and Transformer-based models in various benchmark datasets. However, the architecture is inherently complex, which introduces challenges in model training, hyperparameter tuning, and optimization. Achieving optimal performance often requires specialized hardware configurations and extensive computational resources. This complexity can limit its adoption in general-purpose systems and real-time applications. Despite these challenges, FMViT represents a significant advancement in deep learning architectures, offering a promising direction for future research in high-performance computer vision systems.

The paper *"FMViT: A Multiple-Frequency Mixing Vision Transformer (2023)"* introduces an advanced hybrid architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to improve both accuracy and computational efficiency in visual recognition tasks. Traditional CNNs are highly effective in capturing local spatial features, while Vision Transformers excel in modeling global dependencies. FMViT aims to integrate these complementary strengths into a unified framework. The proposed model incorporates innovative components such as frequency-mixing modules and feature fusion blocks. These mechanisms enable the network to process visual information across multiple frequency domains, effectively capturing both fine-grained details and global contextual information. This design significantly enhances the model's ability to handle complex visual tasks such as object detection, tracking, and scene understanding, making it particularly suitable for applications like retail analytics and surveillance systems. One of the major advantages of FMViT is its ability to achieve high accuracy while maintaining efficient computational performance. The model is designed to reduce redundancy in feature extraction and improve latency, making it viable for deployment in resource-constrained environments. Additionally, its hybrid architecture allows it to outperform traditional CNN-based and Transformer-based models in various benchmark datasets. However, the architecture is inherently complex, which introduces challenges in model training, hyperparameter tuning, and optimization. Achieving optimal performance often requires specialized hardware configurations and extensive computational resources. This complexity can limit its adoption in general-purpose systems and real-time applications. Despite these challenges, FMViT represents a significant advancement in deep learning architectures, offering a promising direction for future research in high-performance computer vision systems.

The paper *"FLORA: Efficient Synthetic Data Generation for Object Detection in Low-Data Regimes via Fine-Tuning Flux LoRA (2025)"* addresses a critical challenge in deep learning—limited availability of labeled training data. High-quality annotated datasets are essential for training robust object detection models, but acquiring such data is often expensive, time-consuming, and labor-intensive. To overcome this limitation, the authors propose a novel approach that leverages Low-Rank Adaptation (LoRA) to fine-tune diffusion models for generating synthetic training data. The FLORA framework creates realistic synthetic images that mimic real-world scenarios, which can then be used to train object detection models. This approach significantly reduces the dependency on large labeled datasets while maintaining strong model performance. One of the key advantages of FLORA is its computational efficiency. By using LoRA-based fine-tuning, the system minimizes training overhead and enables faster adaptation to new tasks. The generated synthetic datasets help improve model generalization, particularly in low-data environments, making it highly valuable for applications such as retail analytics, where labeled data may be scarce. However, the approach has certain limitations. Synthetic data may not fully capture the complexity and variability of real-world environments, leading to potential gaps in model robustness. Additionally, models trained heavily on synthetic data may struggle to adapt to unseen real-world conditions, especially in dynamic and unpredictable scenarios. Despite these challenges, FLORA presents a promising solution for data-scarce applications and highlights the growing importance of synthetic data generation in modern deep learning workflows.

The study "Object Detection Using Convolutional Neural Networks and Transformer-Based Models: A Review (2023)" provides a comprehensive comparative analysis of two dominant paradigms in modern object detection—Convolutional Neural Networks (CNNs) and Transformer-based architectures. CNN-based models, such as Faster R-CNN, YOLO, and SSD, have been widely adopted due to their efficiency in capturing local spatial features and their suitability for real-time applications. On the other hand, Transformer-based models, including DETR (Detection Transformer), introduce attention mechanisms that enable the modeling of long-range dependencies and global contextual relationships within images. The paper reviews multiple state-of-the-art detection frameworks and evaluates their performance using benchmark datasets such as COCO and PASCAL VOC. It highlights that CNN-based models generally offer faster inference and lower computational requirements, making them suitable for real-time applications like surveillance and retail analytics. In contrast, Transformer-based models achieve higher accuracy in complex scenes by effectively capturing global context, although they often require longer training times and higher computational resources. A key contribution of the study is its exploration of hybrid architectures that combine CNN feature extraction with Transformer-based attention mechanisms. These hybrid models aim to balance accuracy and efficiency, addressing the limitations of individual approaches. The paper also discusses challenges such as scalability, data requirements, and model interpretability. However, the study is primarily theoretical and lacks experimental validation or standardized benchmarking across different models. This limits its practical applicability in real-world deployments, where performance evaluation under varying conditions is essential. Despite this limitation, the paper provides valuable insights into the evolution of object detection techniques and highlights future research directions in integrating CNN and Transformer architectures for improved performance.

The paper *"Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment (2024)"* presents a practical implementation of a computer vision-based system designed to analyze customer behavior in retail spaces. The study focuses on leveraging state-of-the-art object detection and tracking algorithms to extract meaningful insights from surveillance video data. Specifically, the system utilizes YOLOv5 for accurate and real-time human detection, combined with DeepSORT for robust multi-object tracking across video frames. The methodology involves detecting customers in each frame and assigning unique identifiers to track their movement throughout the store. The tracked positional data is then aggregated to generate heatmaps, which visually represent areas of high and low customer engagement. These heatmaps provide valuable insights into customer flow patterns, dwell time, and frequently visited zones, enabling retailers to optimize store layouts and product placement strategies. One of the major strengths of this study is its real-time processing capability and scalability, making it suitable for deployment in modern retail environments. The system is non-intrusive, relying solely on video data without requiring additional sensors, thereby reducing implementation complexity. Furthermore, the integration of detection, tracking, and visualization into a unified framework enhances its practical applicability. However, the study also highlights certain limitations. Many retail environments lack the necessary infrastructure, such as high-quality cameras and computational resources, required for effective deployment. Additionally, traditional methods are still preferred in some cases due to their simplicity and lower cost. Despite these challenges, the research demonstrates the

significant potential of computer vision technologies in transforming retail analytics and improving decision-making processes.

The research paper *"Repeatability of Fine-Tuning Large Language Models Illustrated Using QLoRA (2024)"* investigates the consistency and reliability of fine-tuning large language models (LLMs) using the Quantized Low-Rank Adaptation (QLoRA) technique. Fine-tuning large models is a critical step in adapting pre-trained models to specific tasks; however, it often requires significant computational resources and may produce inconsistent results across different training runs. The study evaluates the repeatability of QLoRA by conducting multiple fine-tuning experiments under identical conditions. It utilizes benchmark datasets and compares performance metrics such as accuracy, loss, and generalization capability across different trials. The results demonstrate that QLoRA significantly reduces computational requirements, enabling efficient fine-tuning on single GPUs while maintaining competitive performance levels. This makes it particularly suitable for resource-constrained environments and applications requiring cost-effective deployment. A key contribution of the paper is its focus on reproducibility, which is often overlooked in machine learning research. The findings reveal that while QLoRA offers efficiency and scalability, there is noticeable variability in model performance across different runs. This inconsistency raises concerns about the stability and reliability of fine-tuned models, especially in applications where consistent outputs are critical. The study also discusses factors influencing variability, such as initialization randomness, data sampling, and optimization strategies. While QLoRA presents a promising approach for efficient model adaptation, the lack of consistent reproducibility highlights the need for further research in improving training stability. Overall, the paper provides valuable insights into the trade-offs between efficiency and reliability in modern deep learning workflows.

The paper *"FCOS: Fully Convolutional One-Stage Object Detection (2019)"* introduces an innovative anchor-free object detection framework that simplifies traditional detection pipelines. Conventional object detection models rely on anchor boxes and region proposals, which increase computational complexity and require extensive hyperparameter tuning. FCOS eliminates the need for anchor boxes by treating object detection as a per-pixel prediction problem, similar to semantic segmentation. The proposed method predicts object bounding boxes directly at each pixel location and introduces a "center-ness" branch to improve detection accuracy. This center-ness score helps suppress low-quality predictions and ensures that detections closer to the center of objects are prioritized. By removing anchor boxes, FCOS reduces model complexity and simplifies the training process while maintaining competitive performance compared to anchor-based detectors. One of the key advantages of FCOS is its flexibility and efficiency, as it can easily adapt to different object scales without requiring predefined anchor configurations. The model also integrates well with Feature Pyramid Networks (FPN) to handle multi-scale object detection effectively. Experimental results on benchmark datasets demonstrate that FCOS achieves comparable accuracy to state-of-the-art detectors while offering a simpler and more streamlined architecture. However, the method faces challenges in detecting small or heavily overlapping objects, where precise localization becomes difficult. Additionally, achieving optimal performance requires careful tuning of feature pyramid levels and training parameters. Despite these limitations, FCOS represents a significant step toward simplifying object detection frameworks and has influenced the development of subsequent anchor-free detection models.

The study *"Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows (2021)"* proposes a novel vision backbone that addresses the limitations of traditional Vision Transformers in handling high-resolution images. Standard Transformer models require global self-attention, which leads to high computational complexity, making them less suitable for large-scale visual tasks. The Swin Transformer introduces a hierarchical architecture with shifted window-based self-attention to overcome these challenges. The model divides images into non-overlapping windows and computes self-attention within each window, significantly reducing computational cost. To enable cross-window information exchange, the windows are shifted between layers, allowing the model to capture both local and global contextual relationships. This hierarchical design enables the extraction of multi-scale features, making the model suitable for a wide range of tasks, including image classification, object detection, and semantic segmentation. One of the major strengths of the Swin Transformer is its scalability and efficiency. It achieves state-of-the-art performance on benchmark datasets such as ImageNet and COCO while maintaining manageable computational requirements. The

model's ability to handle multi-scale features effectively makes it particularly useful in complex visual environments, including retail analytics and surveillance systems. However, the architecture is complex and requires substantial computational resources for training. It also depends heavily on large-scale datasets to achieve optimal performance, which may limit its applicability in data-constrained scenarios. Despite these challenges, the Swin Transformer represents a significant advancement in vision models and has paved the way for further research in Transformer-based computer vision architectures.

The paper *"End-to-End Object Detection with Transformers (DETR) (2020)"* introduces a fundamentally new paradigm in object detection by reformulating the problem as a direct set prediction task. Unlike traditional detection frameworks that rely on anchor boxes, region proposals, and post-processing techniques such as non-maximum suppression (NMS), DETR employs a Transformer-based encoder–decoder architecture to predict a fixed set of objects directly from an image. This design significantly simplifies the detection pipeline, reducing the need for hand-crafted components and complex heuristics. The architecture combines a convolutional backbone for feature extraction with a Transformer module that models global relationships across the entire image. The encoder processes image features, while the decoder predicts object bounding boxes and class labels using learned object queries. A bipartite matching loss (Hungarian algorithm) is used during training to uniquely match predicted outputs with ground truth objects, ensuring one-to-one correspondence and eliminating duplicate detections.A major advantage of DETR is its ability to capture long-range dependencies and contextual information, which improves detection performance for large and well-separated objects. Additionally, the end-to-end training framework enhances interpretability and reduces engineering complexity. However, DETR also has notable limitations. It requires significantly longer training times compared to traditional detectors due to the complexity of the Transformer architecture. Furthermore, its performance in detecting small objects is relatively weaker, primarily because of limited feature resolution and slower convergence.Despite these challenges, DETR represents a significant advancement in object detection research and has inspired numerous subsequent models that aim to improve training efficiency and small object detection performance.

The paper *"SSD: Single Shot MultiBox Detector (2016)"* presents a fast and efficient object detection framework designed for real-time applications. Unlike two-stage detectors such as Faster R-CNN, which first generate region proposals and then classify them, SSD performs object localization and classification in a single forward pass of a deep neural network. This design significantly reduces computational overhead and enables high-speed inference. The SSD architecture utilizes a base convolutional network (commonly VGG-16) for feature extraction, followed by multiple convolutional layers that predict bounding boxes and class probabilities at different scales. By leveraging feature maps of varying resolutions, SSD can detect objects of different sizes more effectively. It also employs default bounding boxes (anchors) with multiple aspect ratios, allowing the model to capture diverse object shapes. One of the key strengths of SSD is its ability to achieve high processing speeds while maintaining reasonable accuracy, making it suitable for applications such as video surveillance, autonomous driving, and retail analytics. The model is particularly effective for detecting medium and large-sized objects and can process images at real-time frame rates on standard hardware. However, SSD has certain limitations. Its accuracy is generally lower than that of two-stage detectors, especially in scenarios involving small objects or complex backgrounds. The reliance on fixed anchor boxes can also lead to localization errors and reduced precision in crowded scenes. Additionally, performance can degrade when objects vary significantly in scale or are partially occluded.Despite these drawbacks, SSD remains a widely used baseline for real-time object detection due to its simplicity, efficiency, and balanced trade-off between speed and accuracy.

The paper *"A Survey on LoRA of Large Language Models (2025)"* provides a comprehensive overview of Low-Rank Adaptation (LoRA), an efficient technique for fine-tuning large language models (LLMs). Traditional fine-tuning methods require updating all model parameters, which is computationally expensive and memory-intensive. LoRA addresses this challenge by introducing low-rank decomposition matrices that adapt only a subset of parameters while keeping the original model weights frozen. The study explores various aspects of LoRA, including its application in downstream tasks, cross-task generalization, and efficiency optimization. It highlights that LoRA significantly reduces

the number of trainable parameters, enabling faster training and lower resource consumption without compromising model performance. This makes it particularly suitable for deployment in resource-constrained environments and for applications requiring rapid model adaptation. Another important contribution of the paper is its discussion on data privacy and federated learning. Since LoRA updates only a small portion of the model, it allows decentralized training and reduces the risk of exposing sensitive data. The survey also examines different variants and extensions of LoRA, demonstrating its flexibility across diverse machine learning tasks. However, the study identifies several challenges associated with LoRA. Managing multiple LoRA modules for different tasks can become complex, especially when integrating them into a single system. Additionally, while LoRA improves efficiency, full fine-tuning of large models may still be necessary for achieving optimal performance in certain high-precision applications. Despite these limitations, LoRA represents a significant advancement in efficient model adaptation and is expected to play a key role in the future development of scalable AI systems.

The paper *"Real-Time Object Detection System with YOLO and CNN Models: A Review (2022)"* provides an in-depth analysis of real-time object detection techniques, with a particular focus on the YOLO (You Only Look Once) family of algorithms and Convolutional Neural Networks (CNNs). The study traces the evolution of YOLO from its early versions to more advanced variants, highlighting improvements in detection speed, accuracy, and computational efficiency. YOLO is a single-stage detector that processes the entire image in one pass, dividing it into grids and predicting bounding boxes and class probabilities simultaneously. This approach enables extremely fast inference speeds, with some versions achieving up to 125 frames per second. CNNs play a crucial role in feature extraction, allowing the model to learn hierarchical representations of objects within images.The paper emphasizes the advantages of YOLO-based systems, including real-time performance, scalability, and suitability for applications such as surveillance, traffic monitoring, and retail analytics. It also discusses techniques for improving detection accuracy, such as multi-scale training, anchor box optimization, and feature pyramid networks.However, the study also highlights certain limitations. YOLO models may exhibit higher localization errors compared to two-stage detectors, particularly in detecting small or densely packed objects. Additionally, CNN-based approaches can experience performance degradation in complex scenes with high object density, as overlapping objects may not be accurately distinguished.Despite these challenges, YOLO remains one of the most widely used object detection frameworks due to its balance between speed and accuracy, making it a cornerstone technology in real-time computer vision applications.

The paper *"Real-Time Flying Object Detection with YOLOv8 (2024)"* focuses on developing a robust object detection system tailored for detecting fast-moving and small flying objects in real time. This problem is particularly challenging due to factors such as high object velocity, small object size, motion blur, and frequent occlusions. The study leverages the latest YOLOv8 architecture, which introduces improvements in detection accuracy, speed, and model efficiency compared to earlier YOLO versions. The methodology involves training the YOLOv8 model on a large and diverse dataset, followed by fine-tuning using transfer learning on domain-specific real-world data. This approach enables the model to generalize effectively while adapting to specific application requirements. The system achieves a mean average precision (mAP) of approximately 83.5% and operates at around 50 frames per second, demonstrating its suitability for real-time applications such as surveillance, drone monitoring, and security systems. A key strength of the study is its focus on handling challenging detection scenarios, including small object detection and occlusion. The model incorporates advanced techniques such as improved feature extraction, better anchor-free detection strategies, and optimized training pipelines to enhance performance. However, the study also identifies limitations. Detecting small and partially occluded objects remains challenging, especially in cluttered environments. Additionally, the lack of detailed official documentation for the YOLOv8 architecture can create difficulties for researchers and developers attempting to implement or modify the model. Despite these challenges, the paper demonstrates the effectiveness of YOLOv8 as a powerful tool for real-time object detection and highlights its potential for various advanced computer vision applications.

The study *"Focal Loss for Dense Object Detection (RetinaNet) (2018)"* addresses a critical challenge in one-stage object detection models—class imbalance between foreground objects and background regions. Traditional one-stage

detectors often struggle with this imbalance, as the overwhelming number of easy negative samples dominates the training process, leading to suboptimal performance. To overcome this issue, the authors introduce Focal Loss, a modified version of the standard cross-entropy loss function that dynamically down-weights well-classified examples and focuses training on hard, misclassified samples. This innovation enables one-stage detectors like RetinaNet to achieve accuracy levels comparable to two-stage detectors such as Faster R-CNN, while maintaining the inherent speed advantage of single-stage architectures. The RetinaNet model integrates a Feature Pyramid Network (FPN) backbone, allowing it to detect objects at multiple scales effectively, which is particularly beneficial in dense detection scenarios such as crowd analysis and retail environments. The application of Focal Loss significantly improves detection performance in situations with high object density and class imbalance, making it suitable for real-world applications involving surveillance and customer analytics. However, despite its advantages, the method has limitations. It may still struggle in cases of extreme imbalance where rare classes are significantly underrepresented, and its performance can degrade when applied to sparse datasets with limited variability. Additionally, tuning the focusing parameter of the loss function requires careful experimentation to achieve optimal results. Overall, the introduction of Focal Loss represents a major advancement in object detection research, providing a practical solution to one of the key limitations of one-stage detectors.

The paper *"Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment (2024)"* presents a practical and scalable system for analyzing customer behavior within retail spaces using computer vision techniques. The system leverages YOLOv5 for real-time human detection due to its high accuracy and speed, while DeepSORT is employed for robust multi-object tracking, enabling consistent identification of individuals across video frames. By combining detection and tracking, the system captures detailed movement trajectories of customers throughout the store. These trajectories are aggregated to generate heatmaps, which visually represent areas of high and low customer engagement. Such visualizations provide valuable insights into customer flow patterns, dwell time, and frequently visited zones, enabling retailers to optimize store layout, product placement, and marketing strategies. One of the key strengths of this approach is its non-intrusive nature, as it relies solely on existing surveillance infrastructure without requiring additional sensors. Furthermore, the system supports real-time analytics, making it suitable for dynamic retail environments where immediate insights are crucial. However, the approach also faces several challenges. Its performance is highly dependent on camera quality, positioning, and lighting conditions, which can significantly affect detection and tracking accuracy. In crowded environments, occlusions and overlapping individuals can lead to tracking errors and reduced reliability. Additionally, the use of surveillance-based analytics raises privacy concerns, which may impact user acceptance and regulatory compliance. Despite these limitations, the study demonstrates the strong potential of integrating detection and tracking models for intelligent retail analytics.

The research *"Object Detection and Visual Intelligence in Retail Environments: A Deep Learning Approach for Inventory and Behavior Analytics"* proposes a comprehensive framework that integrates object detection and behavioral analytics to provide a holistic understanding of retail operations. The system utilizes advanced deep learning models such as YOLOv8 for real-time object detection and Mask R-CNN for instance segmentation, enabling precise identification of both customers and products within the retail environment. By combining these models with scene-aware post-processing techniques, the framework can analyze customer interactions with products, track movement patterns, and monitor inventory levels simultaneously. This dual capability makes the system particularly valuable for applications such as shelf monitoring, demand forecasting, and customer engagement analysis. The architecture is designed to support real-time analytics, allowing retailers to make immediate decisions based on current store conditions. Additionally, the system is scalable and can be deployed across multiple store locations, making it suitable for large retail chains. However, the approach requires a substantial amount of labeled training data to achieve high accuracy, which can be costly and time-consuming to obtain. The system is also sensitive to occlusions, complex store layouts, and varying lighting conditions, which can affect detection performance and reliability. Furthermore, integrating multiple deep learning models increases computational complexity, requiring high-performance hardware for efficient operation. Despite these challenges, the study highlights the potential of combining object detection and behavioral analytics to create intelligent retail systems capable of delivering comprehensive insights.

The paper *"Analyzing Customer Behavior In-Store: A Review of Available Technologies"* provides a systematic overview of the various technologies used to analyze customer behavior in physical retail environments. The study categorizes different measurement objectives, such as person detection, product interaction, and movement path analysis, and examines the corresponding sensing technologies used to achieve these objectives. These technologies include camera-based systems, RFID tracking, Bluetooth beacons, Wi-Fi analytics, and sensor-based solutions. The paper also introduces a suitability matrix that helps retailers select appropriate technologies based on factors such as cost, accuracy, scalability, and privacy considerations. One of the key contributions of the study is its comprehensive comparison of different approaches, highlighting the strengths and limitations of each technology. For instance, camera-based systems provide rich visual data and enable detailed behavioral analysis, while sensor-based methods offer simplicity and lower implementation costs. The study emphasizes the importance of selecting the right combination of technologies to achieve optimal results in specific retail scenarios. However, the paper is primarily a review and does not propose any new algorithms or methodologies. It also lacks detailed implementation analysis and real-world validation, limiting its direct applicability for system development. Additionally, the rapid evolution of AI-based techniques may render some of the reviewed technologies less relevant over time. Despite these limitations, the study serves as a valuable reference for understanding the landscape of customer behavior analysis technologies and provides a foundation for future research in intelligent retail systems.

The study *"A Customer Behavior Recognition Method for Flexibly Adapting to Target Changes in Retail Stores"* proposes an alternative approach to behavior recognition that focuses on adaptability rather than relying solely on deep learning models. Traditional behavior recognition systems often require extensive retraining when customer behavior patterns change, which can be time-consuming and computationally expensive. To address this issue, the proposed method decomposes complex customer behaviors into simpler components, referred to as "behavior primitives," such as object motion, spatial relationships, and interaction sequences. These primitives are then combined using pattern-matching techniques to identify higher-level behaviors. This modular approach allows the system to adapt to new behavior patterns without requiring complete retraining, making it more flexible and efficient in dynamic retail environments. One of the key advantages of this method is its ability to handle evolving customer behaviors with minimal computational overhead. It also provides a more interpretable framework compared to deep learning models, as the behavior recognition process is based on explicit rules and patterns. However, the method has several limitations. Its accuracy is generally lower than that of deep learning-based approaches, particularly in complex or highly dynamic scenarios where behaviors are difficult to decompose into simple primitives. Additionally, the reliance on predefined patterns may limit its ability to generalize to unseen behaviors. Despite these challenges, the study presents an innovative perspective on behavior recognition and highlights the importance of flexibility and adaptability in retail analytics systems.

The paper *"ByteTrack: Multi-Object Tracking by Associating Every Detection Box"* introduces an advanced multi-object tracking (MOT) approach that significantly improves tracking performance by utilizing all detection boxes, including those with low confidence scores. Traditional tracking methods often discard low-confidence detections, which can result in missed objects and fragmented trajectories, especially in crowded environments. ByteTrack addresses this limitation by employing a two-stage data association strategy, where high-confidence detections are matched first, followed by the association of low-confidence detections with existing tracks. This approach enhances object continuity and improves recall without compromising precision. The method integrates seamlessly with modern object detectors such as YOLO, making it suitable for real-time applications like surveillance and retail analytics. Experimental evaluations on benchmark datasets demonstrate that ByteTrack achieves state-of-the-art performance in terms of tracking accuracy and identity preservation. One of its key strengths lies in its ability to maintain stable tracking in challenging scenarios involving occlusions, fast movement, and dense crowds. However, the approach is highly dependent on the quality of the initial object detection stage; noisy or inaccurate detections can propagate errors into the tracking process. Additionally, incorporating low-confidence detections increases computational complexity, particularly in densely populated scenes, which may impact real-time performance on limited hardware. Despite these challenges, ByteTrack represents a significant advancement in multi-object tracking, offering a practical solution for improving tracking robustness and accuracy in dynamic environments.

The research paper *"YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information"* introduces a novel object detection framework that enhances feature learning through the use of programmable gradient information. Unlike traditional models that rely on fixed backpropagation mechanisms, YOLOv9 allows selective control over gradient flow, enabling the model to prioritize learning from more relevant features while suppressing less informative ones. This adaptive learning strategy improves detection accuracy and generalization, particularly in complex visual environments where distinguishing between objects and background noise is challenging. The architecture builds upon the strengths of previous YOLO versions by maintaining real-time performance while incorporating advanced optimization techniques to refine feature extraction and representation. One of the key advantages of YOLOv9 is its flexibility, as it can be tailored to specific tasks by adjusting gradient propagation rules, making it suitable for diverse applications such as surveillance, autonomous systems, and retail analytics. The model demonstrates improved performance in detecting objects under varying conditions, including occlusions and scale variations. However, the introduction of programmable gradients increases the complexity of the model, making it more difficult to implement and tune effectively. Training the model requires significant computational resources, particularly GPUs, which may limit its accessibility for smaller-scale applications. Additionally, the lack of extensive public benchmarking results makes it challenging to evaluate its performance consistently across different datasets and use cases. Despite these limitations, YOLOv9 represents an innovative step forward in object detection, offering a new perspective on adaptive learning in deep neural networks.

The paper *"Feature Pyramid Networks for Object Detection"* presents a powerful approach to improving object detection across multiple scales by constructing a hierarchical feature representation within convolutional neural networks. Traditional detection models often struggle with scale variation, particularly when detecting small objects, as deeper layers capture high-level semantic information but lose spatial resolution. Feature Pyramid Networks (FPN) address this issue by introducing a top-down architecture with lateral connections, which combine high-level semantic features with lower-level spatial details. This results in a multi-scale feature pyramid that enhances the model's ability to detect objects of varying sizes effectively. FPN integrates seamlessly with existing detection frameworks such as Faster R-CNN and RetinaNet, significantly improving their performance without requiring substantial modifications to the underlying architecture. One of the key strengths of FPN is its ability to improve detection accuracy for small and medium-sized objects, which is particularly important in applications like retail analytics and surveillance. The method also enhances feature reuse, making it computationally efficient compared to constructing image pyramids explicitly. However, the integration of multi-scale feature fusion introduces additional computational overhead, which can impact inference speed, especially in real-time applications. Furthermore, despite its improvements, FPN still faces challenges in detecting extremely small or heavily occluded objects in complex environments. Nonetheless, the concept of feature pyramids has become a foundational component in modern object detection architectures, influencing the design of numerous state-of-the-art models.

The study *"Enhancing Facial Recognition Accuracy in Low-Light Conditions Using Convolutional Neural Networks"* focuses on addressing one of the major challenges in computer vision—accurate facial recognition under poor lighting conditions. In real-world scenarios such as surveillance and retail environments, lighting conditions can vary significantly, leading to degraded image quality and reduced recognition accuracy. The proposed approach combines CNN-based image enhancement techniques with deep learning-based facial recognition models to improve feature extraction from low-light images. The system first applies preprocessing methods, such as noise reduction and contrast enhancement, to improve image visibility. These enhanced images are then fed into a CNN-based recognition model that extracts robust facial features, enabling accurate identification even under challenging lighting conditions. The study demonstrates that this combined approach significantly improves recognition performance compared to traditional methods, particularly in dimly lit environments. One of the key strengths of the method is its ability to enhance feature representation without requiring significant changes to the underlying recognition model. However, the approach also has limitations. It requires high-quality and diverse training data to generalize effectively across different lighting conditions. Additionally, the model may struggle in highly dynamic environments where lighting conditions change rapidly or where faces are partially occluded. The computational overhead introduced by image enhancement processes can also impact real-time performance. Despite these challenges, the study provides valuable insights into improving the robustness of facial recognition systems and highlights the importance of preprocessing techniques in enhancing model performance.

The paper *"TrackFormer: Multi-Object Tracking with Transformers"* introduces a unified framework that combines object detection and tracking using Transformer-based architectures. Unlike traditional tracking systems that rely on separate detection and association stages, TrackFormer integrates both tasks into a single end-to-end model. The approach leverages the self-attention mechanism of Transformers to model temporal dependencies across video frames, enabling the system to associate object detections over time effectively. By using object queries that persist across frames, the model can maintain consistent identities for tracked objects, resulting in improved tracking accuracy and reduced identity switches. This unified architecture simplifies the tracking pipeline and eliminates the need for complex post-processing steps, making it conceptually elegant and easier to maintain. The model demonstrates strong performance on benchmark datasets, achieving high accuracy in multi-object tracking tasks, particularly in scenarios with complex motion and occlusions. However, the use of Transformer architectures introduces significant computational demands, requiring powerful hardware and large annotated datasets for training. This can limit its applicability in real-time or resource-constrained environments such as small-scale retail setups. Additionally, the complexity of the model makes it challenging to optimize and deploy efficiently. Despite these limitations, TrackFormer represents a significant advancement in tracking methodologies, showcasing the potential of Transformer-based approaches in achieving highly accurate and temporally consistent object tracking.

3. CONCLUSIONS

The literature survey highlights the rapid advancements in computer vision, deep learning, and artificial intelligence techniques for retail analytics and intelligent monitoring systems. A wide range of approaches, including Convolutional Neural Networks (CNNs), Transformer-based architectures, hybrid models, and advanced tracking algorithms, have been explored to improve object detection, multi-object tracking, and behavioral analysis. Techniques such as YOLO, SSD, RetinaNet, DETR, and Vision Transformers demonstrate significant progress in achieving a balance between accuracy and real-time performance, while tracking methods like DeepSORT, ByteTrack, and TrackFormer enhance the ability to maintain consistent object identities across video frames. Additionally, supporting technologies such as Feature Pyramid Networks, Focal Loss, and synthetic data generation methods have contributed to improving detection performance under challenging conditions such as scale variation, occlusion, and data scarcity.

The survey also emphasizes the growing importance of retail-specific applications, including customer behavior analysis, heatmap generation, demographic estimation, and store layout optimization. Many studies demonstrate the effectiveness of integrating detection and tracking models to extract actionable insights such as footfall, dwell time, and engagement patterns. However, several limitations persist, including dependency on high-quality data, sensitivity to environmental conditions, computational complexity, and challenges in real-time deployment. Privacy concerns and infrastructure requirements further impact the adoption of such systems in practical retail environments.

Despite these challenges, the reviewed works collectively indicate that AI-driven retail analytics systems have strong potential to transform traditional retail operations into intelligent, data-driven ecosystems. The integration of real-time video analytics with advanced machine learning models enables more accurate, scalable, and automated decision-making processes. Overall, this survey establishes a strong foundation for the development of the proposed Retail Insight Generator, which aims to address existing limitations by providing a robust, efficient, and practical solution for real-time customer behavior analysis and retail optimization.

4. REFERENCES

[1]. When AI Meets Store Layout Design: A Review - S. Kumar et al. - 2022.

[2]. Deep Learning Based Approach to Detect Customer Age, Gender and Expression in Surveillance Video - R. Rothe et al. - 2020.

[3]. FMViT: A Multiple-Frequency Mixing Vision Transformer - Y. Li et al. - 2023.

[4]. Analyzing Computer Vision Models for Detecting Customers: A Practical Experience in a Mexican Retail - J. Pérez et al. - 2024.

[5]. FLORA: Efficient Synthetic Data Generation for Object Detection in Low-Data Regimes via Fine-Tuning Flux LoRA - H. Zhang et al. - 2025.

[6]. Object Detection Using Convolutional Neural Networks and Transformer-Based Models: A Review - A. Sharma et al. - 2023.

[7]. Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment - M. Khan et al. - 2024.

[8]. Repeatability of Fine-Tuning Large Language Models Illustrated Using QLoRA - T. Dettmers et al. - 2024.

[9]. FCOS: Fully Convolutional One-Stage Object Detection - Z. Tian et al. - 2019.

[10]. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows - Z. Liu et al. - 2021.

[11]. End-to-End Object Detection with Transformers (DETR) - N. Carion et al. - 2020.

[12]. SSD: Single Shot MultiBox Detector - W. Liu et al. - 2016.

[13]. A Survey on LoRA of Large Language Models - Q. Hu et al. - 2025.

[14]. Real-Time Object Detection System with YOLO and CNN Models: A Review - S. Redmon et al. - 2022.

[15]. Real-Time Flying Object Detection with YOLOv8 - A. Jocher et al. - 2024.

[16]. Focal Loss for Dense Object Detection (RetinaNet) - T.-Y. Lin et al. - 2018.

[17]. Advanced Customer Behavior Tracking and Heatmap Analysis with YOLOv5 and DeepSORT in Retail Environment - M. Khan et al. - 2024.

[18]. Analyzing Customer Behavior In-Store: A Review of Available Technologies - L. Grewal et al. - 2021.

[19]. Object Detection and Visual Intelligence in Retail Environments: A Deep Learning Approach for Inventory and Behavior Analytics - P. Sharma et al. - 2023.

[20]. A Customer Behavior Recognition Method for Flexibly Adapting to Target Changes in Retail Stores - Y. Sato et al. - 2020.

[21]. ByteTrack: Multi-Object Tracking by Associating Every Detection Box - Y. Zhang et al. - 2022.

[22]. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information - C.-Y. Wang et al. - 2024.

[23]. Feature Pyramid Networks for Object Detection - T.-Y. Lin et al. - 2017.

[24]. Enhancing Facial Recognition Accuracy in Low-Light Conditions Using Convolutional Neural Networks - S. Gupta et al. - 2021.

[25]. TrackFormer: Multi-Object Tracking with Transformers - T. Meinhardt et al. - 2022.