HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Hybrid Edge-Server Person Re-ID Scalable Microservices with Metadata-Enhanced Features

Duong Minh Quan
quandm.210710@sis.hust.edu.vn

Major: Data Science and Artificial Intelligence

Supervisor: Dr. Dang Tuan Linh _____

Signature

Department: Computer Engineering

School: School of Information and Communications Technology

HANOI, 07/2025

ACKNOWLEDGMENT

Firstly, I would like to express my gratitude to Dr. Dang Tuan Linh for all of his assistance, not just in helping me finish this thesis but also during my four years of university study. I have learned so much from him and those values are extremely valuable to me, which have greatly assisted me in both work and study. Additionally, I also want to sincerely thank Mr. Hung and Ms. Quynh, without whom it's likely I would not have been able to successfully complete this thesis. I am also grateful to other friends, seniors, and contributors, whether in small or large ways, tangible or intangible, for their contributions to this completed thesis today. Finally, I would like to thank my family for always supporting me unconditionally.

ABSTRACT

In today's fast-paced world, effective management and analysis are crucial for maintaining security and enhancing business productivity, particularly as the number of enterprises and the size of enterprises rise. According to a recent report from the General Statistics Office of Vietnam, between 2016 and 2019, there was an average annual increase of 9.8% in the number of enterprises, which is higher than the average annual growth rate of 8.1% observed between 2011 and 2015. Moreover, the behavior of employees can be difficult to manage, and employers can increase productivity if they have valuable information from human behavior. Therefore, it is necessary to implement modern technology, specifically AI, into monitoring systems in order to reduce costs and increase productivity.

However, the majority of current AI implementations rely on centralized servers, making scaling difficult. This thesis proposes a novel AI module that can be installed on edge devices, as a means of overcoming this obstacle. Human detection, human tracking, and human feature extraction are the three primary components of the proposed module. All of these components are directly executed on edge devices. This AI module can be utilized to monitor individuals and collect data that can be used to enhance the productivity of businesses.

I aim to achieve efficient human management and analysis by implementing AI models on edge devices that are readily scalable. The algorithms utilized in this thesis have been successfully implemented on Jetson Nano devices with low computational capability while maintaining above 10 FPS for less than seven people in a single frame. A prototype of this module has been put into practical use and examined in room 405, B1 building at Hanoi University of Science and Technology. The proposed module has the potential to revolutionize office human resource management and analysis, thereby enhancing office security and productivity while reducing costs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AI	Artificial Intelligence
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
EER	Equal Error Rate
FPS	Frames Per Second
GPU	Graphics Processing Unit
ID	Identification
IoT	Internet of Things
IoU	Intersection over Union
mAP	mean Average Precision
NMS	Non-maximum Suppression
Re-ID	Re-identification
ROI	Region Of Interest
USB	Universal Serial Bus

CHAPTER 1. INTRODUCTION

This chapter provides the understanding of the current circumstances that SMEs are facing due to their limited financial ability to implement solutions for enhancing user experience, particularly using AI solutions. This also forms the foundation that brought me to develop this thesis, helping SMEs access person Re-ID technologies with limited budgets.

1.1 Background and Motivation

Small and Medium Enterprises (SMEs) across various industries are facing an extraordinary challenge in today's competitive market. Customer expectations have fundamentally shifted from simple product transactions to demanding rich, personalized experiences. This change is particularly evident in sectors like Food & Beverage (F&B) and retail, where 65% of customers report that positive experiences influence their purchasing decisions more than traditional advertising [1].

SMEs operate under much tighter financial constraints than large corporations. In the F&B sector alone, 45% of businesses report that raw materials account for over 30% of their selling prices, leaving little room for major technology investments [2]. Over 60% of F&B businesses have experienced revenue decreases while facing rising operational costs including rent, labor, and materials [3].

This creates an "innovation deadlock" where SMEs:

- Recognize the critical need for better customer experience solutions.
- Understand that technology could provide competitive advantages.

Modern AI technologies like Re-ID offer powerful solutions for understanding customer behavior, optimizing store layouts, and creating personalized experiences. Re-ID systems can seamlessly track customer movements across different areas of a store, measure how long customers spend in specific sections, and identify popular pathways and bottlenecks. This technology enables businesses to provide tailored assistance, highlight relevant promotions based on customer interests, and optimize staff allocation in real-time.

However, traditional Re-ID systems present significant economic barriers that make them inaccessible to most SMEs. Conventional implementations require expensive GPU-powered edge devices. When scaled across multiple cameras needed for comprehensive coverage, these costs become prohibitive.

The high computational requirements of traditional Re-ID systems also demand

powerful central servers for data processing and storage, further inflating the total cost of ownership. For SMEs already struggling with thin profit margins, these substantial upfront investments often exceed their entire annual technology budgets.

Therefore, there is an urgent need for affordable, scalable Re-ID solutions designed specifically for SME use. Such systems should lower hardware costs by using efficient CPU-based processing at the edge, reduce complex setup requirements, and provide useful customer experience improvements that help SMEs compete on service quality rather than just price. By making intelligent customer interaction technologies accessible to all businesses, we can help companies of all sizes improve customer satisfaction, build loyalty, and achieve steady growth in today's experience-focused marketplace.

1.2 Challenges and Current Solutions

The architectural design of a person Re-ID system is a critical consideration, with each approach presenting distinct trade-offs. The chosen architecture directly impacts the system's cost, scalability, and real-time performance, making it an especially important factor for SMEs operating under budget constraints. The primary architectural models are centralized, edge-based, and hybrid, yet none of these fully address the unique challenges faced by small businesses.

1.2.1 Fully Centralized Architecture

The conventional approach involves a centralized architecture where standard cameras transmit their video feeds over a network to a single, powerful server. This central server is responsible for all computationally demanding tasks, such as person detection, feature extraction, and identity matching.

While this model simplifies on-site hardware, it introduces significant drawbacks that render it impractical for most SMEs:

- **Network Dependency and Costs:** The continuous streaming of video from multiple cameras requires substantial network bandwidth, leading to high operational costs. The system's reliability is also contingent on network stability, as latency or packet loss can result in incomplete data.
- **High Server Costs:** The computational load on the central server scales with the number of cameras, necessitating a significant upfront investment in high-performance server hardware. For many SMEs, this cost is prohibitive. This architecture also creates a single point of failure.
- **Management Complexity:** A Re-ID pipeline consists of multiple processing stages. Managing this complex workflow for numerous concurrent video streams

on a single machine presents a considerable technical challenge.

1.2.2 Fully Edge-Based Architecture

To mitigate the network dependencies of the centralized model, an edge-based architecture places computational power on devices located near the cameras. These "edge" devices process video locally and transmit only lightweight metadata, such as feature vectors, to a central location.

This design reduces network bandwidth requirements and enhances resilience to network disruptions. However, it introduces its own distinct disadvantages for SMEs:

- **High Cumulative Hardware Cost:** The primary issue is the cost of the edge devices. While a single unit may be affordable, the expense escalates with each camera added, making large-scale deployments costly.
- **Distributed Maintenance:** Managing a distributed fleet of edge devices is operationally more complex than maintaining a single server, increasing the burden of software updates and hardware troubleshooting.

1.2.3 Hybrid Architecture

A hybrid architecture attempts to strike a balance by distributing the workload between edge devices and a central server. For instance, a low-cost edge device might handle initial person detection, while the more intensive matching tasks are offloaded to the server.

This approach aims to reduce hardware costs at the edge, but it presents its own set of complexities:

- **System Integration Challenges:** Dividing tasks creates a more intricate, multi-tiered system. Ensuring seamless communication and efficient integration between the edge and server components is a significant engineering task.
- **Potential for Latency:** The handoff of data between the edge and the server can introduce processing delays. In applications requiring immediate responses, this latency can undermine the system's utility.
- **Workload Balancing:** Achieving an optimal balance is difficult. If the edge device is underpowered, it can become a bottleneck. Conversely, if too much processing is offloaded to the server, the architecture reintroduces the bandwidth and cost issues of the centralized model.

In summary, current Re-ID architectures do not present ideal solutions for SMEs due to challenges related to cost, network dependency, and complexity. This "innovation

deadlock" hinders smaller businesses from adopting this valuable technology. This thesis proposes a novel hybrid solution engineered to be affordable, scalable, and manageable within an SME context.

1.3 Objectives and scope of Thesis

This thesis aims to address the challenges faced by SMEs in adopting person Re-ID technology by developing a cost-effective, scalable solution that combines the benefits of edge computing with centralized processing power.

The primary objective is to design and implement an end-to-end person Re-ID pipeline that can operate efficiently on CPU-based edge devices without requiring expensive GPU acceleration. This approach significantly reduces hardware costs while maintaining acceptable performance levels for real-world applications.

The scope of this thesis encompasses several key areas:

- **System Architecture Design:** Development of a hybrid edge-server architecture that balances computational load between lightweight edge devices and a central server. This design minimizes network bandwidth requirements while keeping hardware costs manageable for SMEs.
- **AI Model Development and Training:** Implementation of custom lightweight models optimized for CPU inference, including:
  - A person detection model achieving 85% mAP at 14 FPS on only 1 CPU core and 512 MB of RAM
  - A gender classification model with 95% accuracy for metadata enhancement
- **Intelligent Retrieval Optimization:** Development of a metadata-enhanced search algorithm that uses gender classification results to reduce the search space during identity matching, improving retrieval speed by 40% and accuracy by 8% compared to traditional approaches.
- **Containerized Deployment:** Implementation of the entire system using containerization technology to ensure easy deployment, consistent performance across different environments, and simplified maintenance procedures.
- **Microservices Implementation:** Breaking down the Re-ID pipeline into independent microservices, particularly for AI model serving through HTTP APIs. This approach enables horizontal scaling, improves system reliability, and allows for independent updates of system components.
- **High-Throughput Model Serving:** Optimization of the inference pipeline to achieve up to 170 requests per second (RPS) for a 7 million parameter model.

This includes full GPU utilization on the central server and elimination of bottlenecks that could limit processing throughput.

- **System Reliability:** Integration of monitoring, alerting, and health check mechanisms to ensure high system availability (99.5% uptime) and quick identification of potential issues.

The ultimate goal is to provide SMEs with an accessible path to implement person Re-ID technology that fits within their budget constraints while delivering the customer experience improvements they need to remain competitive in today's market.

1.4 Contributions

1. An application is deployed on hybrid edge-server devices and uses a microservices architecture, allowing for easy system scaling (increasing the number of cameras). It includes:
    - A custom-trained, lightweight human detection model specifically designed for CPU-based, resource-constrained edge devices.
    - A custom-trained, lightweight gender classification model with 95% accuracy for metadata enhancement.
    - A vector database optimization algorithm for efficient identity retrieval. This uses a person's metadata (gender) to reduce the search space, improving retrieval speed and accuracy.
    - This thesis also provides an interactive web application. It lets users monitor the system, view live camera streams, and search for people using their metadata.

1.5 Organization of Thesis

CHAPTER 2. LITERATURE REVIEW

Chapter 1 established the foundational context by identifying the innovation deadlock faced by SMEs, defining the research aims for creating affordable person Re-ID systems, examining current architectural limitations, and outlining the novel contributions of this thesis. There are two main parts in this chapter. They are presentations about (i) related works in Section 2.1, and (ii) the foundation theory in Section 2.2. Specifically, in Section 2.2, models, frameworks, utilities and algorithms contributing to the making of the end-to-end pipeline will be introduced in detail.

To achieve the thesis objectives, four key technical components will be examined: (i) lightweight object detection using YOLOv11 for CPU-based edge deployment in Section 2.2.1, (ii) efficient object tracking through ByteTrack algorithms in Section 2.2.2, (iii) optimized feature extraction methodologies for resource-constrained environments in Section 2.2.3, light-weight image classification task specific for human's metadata (gender) in section 2.2.4, and (iv) distributed system infrastructure including message queuing in Section 2.2.5, containerization in Section 2.2.7, and vector database optimization in Section 2.2.8.

2.1 Related works

2.1.1 Person Re-Identification

Recent advances in person Re-ID have significantly enhanced performance across various scenario, primarily relying on powerful computational resources. In the context of visible–infrared ReID, Guo et al. (2025) introduced the Region-based Augmentation and Cross Modality Attention (RACA) model [4], which leverages region-level augmentation (PedMix) and a modality feature transfer (MFT) module with cross-attention to reduce interference between modalities, yielding notable improvements on SYSU-MM01 and RegDB benchmarks.

Addressing unsupervised learning, Qin et al. (2025) proposed Attention-based Hybrid Contrastive Learning (AHCL) [5]. Their framework integrates spatial and channel attention with a hybrid contrastive loss, combining cluster-level and instance-level representations to bolster ReID accuracy without labels.

Multimodal ReID has been further advanced by Yan et al. (2025) through FusionSegReID [6], which fuses image features, textual descriptions, and segmentation masks to enhance robustness—especially in occluded or low-quality scenarios. Interactive, language-driven retrieval was pushed forward by Niu et al. (2025) in ChatReID [7]. This framework uses a Vision–Language Model (LVLM)

with Hierarchical Progressive Tuning, enabling interactive, VQA-style queries to improve identity-level matching performance.

Despite the impressive progress of person ReID, most state-of-the-art models demand substantial hardware and computational resources, limiting their practicality on lightweight or embedded devices. To address this, **OSNet** (Omni-Scale Network) was proposed by Zhou et al. [8]. OSNet is a compact yet powerful architecture, specifically designed for efficient deployment. It employs *omni-scale feature learning*, utilizing multiple convolutional streams with varying receptive field sizes within each residual block to capture both fine-grained details and global features. Additionally, OSNet integrates a *Unified Aggregation Gate (UAG)* that adaptively fuses multi-scale features via channel-wise weighting. By leveraging depthwise-separable

convolutions, OSNet significantly reduces computational cost and model size, achieving state-of-the-art performance despite being substantially lighter than standard models like ResNet-50.

Building upon OSNet's efficiency, **LightMBN** (Lightweight Multi-Branch Network) introduced by Herzog et al. [9] further optimizes this backbone architecture. LightMBN expands OSNet by adding specialized global, part-based, and channel-wise branches, thereby enriching feature representations without significant complexity increases. It also incorporates enhanced training techniques, including label smoothing, random erasing, and cosine learning rate schedules, leading to improved generalization performance. Consequently, LightMBN achieves impressive accuracies on widely-used benchmarks like Market-1501 and CUHK03, outperforming many heavier architectures while remaining lightweight and suitable for resource-constrained deployments.

Given their complementary strengths in efficiency and performance, OSNet and LightMBN form a robust backbone choice for deployment in lightweight Re-ID pipelines.

2.1.2 Edge Computing in AI

Edge computing has emerged as a transformative approach in artificial intelligence, bringing computational resources closer to where data is generated and directly to end users. The global edge AI market size was valued at approximately USD 20.78 billion in 2024 and is expected to grow significantly, at a rate of 21.7% annually, from 2025 to 2030 [10]. This growth indicates a strong demand for real-time processing, lower latency, and improved privacy in various AI applications.

Within retail and customer experience contexts, edge computing provides substantial

benefits. It enables retail IT teams to manage cloud expenses effectively by strategically selecting which data to send to the cloud, processing only critical information rather than all raw data [11]. This selective processing is particularly beneficial for person Re-ID systems, where enormous video data streams can be filtered and analyzed locally, and only essential features are transmitted to centralized servers.

Recent research has also concentrated on adapting AI models specifically for edge environments. Novel person Re-ID methods incorporate pedestrian edge features directly into their representations and leverage these edge characteristics to enhance global context feature extraction [12]. Such methods highlight the practical feasibility of deploying sophisticated Re-ID algorithms on devices with limited computational capabilities.

Edge AI is becoming more widely available across many applications. Edge Intelligence, or Edge AI, means moving AI processing from cloud systems directly to edge devices where data is created. This change is important for making AI more available and affordable, especially for small and medium-sized businesses that may find cloud-based AI solutions too expensive [13].

2.1.3 Microservices and Distributed Systems

Microservices have become increasingly popular for building scalable AI systems. In general, microservices focus on modularity, meaning each service handles a specific function independently. These services are loosely connected, easy to deploy individually, and can scale separately [14].

In person Re-ID systems, microservices offer clear advantages, especially when deployed across distributed environments. For example, using microservices in combination with AI-powered edge computing gateways helps handle privacy concerns while efficiently identifying the same person from multiple camera angles and locations [15].

However, using distributed setups in Re-ID introduces new challenges. Traditional Re-ID algorithms typically prioritize accuracy but aren't designed with distributed deployment in mind. Distributed environments demand algorithms that are lightweight and computationally efficient [16]. Thus, there is a significant need for simpler, more efficient Re-ID algorithms suitable for running on multiple edge nodes.

Recent studies have explored distributed frameworks for deep learning applications, particularly using microservices for object detection tasks on edge devices. These frameworks effectively analyze images and videos to extract relevant object information and locations, providing a solid foundation for scalable Re-ID applications involving

many cameras or geographic areas [17].

Cloud-based solutions have also been investigated. Video-based person Re-ID using distributed cloud computing stores pedestrian data and model parameters across multiple cloud servers to improve reliability and reduce failures [18]. However, ongoing cloud service costs can be prohibitive for small and medium-sized businesses.

Introducing a message broker into a microservices architecture provides additional advantages. Message brokers enable seamless communication among microservices by handling data exchange, enhancing reliability, and simplifying integration. Specifically, they help Re-ID systems quickly share person-related data across various cameras and processing units, ensuring low latency and better system scalability.

Overall, the shift toward microservices architecture, combined with the use of message brokers, reflects a broader trend towards modular, scalable, and efficient AI system deployment, making it particularly beneficial for distributed person Re-ID solutions in retail and similar settings [19].

2.2 Foundation theory

2.2.1 Object detection

Object detection technology finds applications across numerous domains including automatic traffic violation systems, identification of unfamiliar persons, digital attendance systems, and autonomous robotic vehicles. The advent of deep learning has dramatically enhanced object detection capabilities. Region-Based Convolutional Neural Networks (R-CNN) [20] represented one of the pioneering breakthroughs in this area, combining CNN architectures [21] with region proposal mechanisms to achieve accurate object localization and classification in images. Subsequent iterations, Fast R-CNN [22] and Faster R-CNN [23], were developed to enhance both processing speed and detection precision compared to the original model. Despite these improvements in detection performance, the multi-step processing pipeline made these approaches impractical for real-time applications.

Modern frameworks such as Detectron2 [24] and EfficientDet [25] have pushed object detection forward considerably. Detectron2 offers flexible deployment of high-performing models but demands extensive setup and typically involves computationally heavy architectures, making it unsuitable for real-time or resource-limited environments. EfficientDet provides a more practical option for devices with constrained resources, though it may struggle to meet strict real-time performance criteria.

You Only Look Once (YOLO) addresses the limitations of multi-stage detection

approaches by reformulating object detection as a single regression problem. YOLO processes the entire image in one forward pass, directly predicting bounding boxes and class probabilities from full images. This unified architecture enables real-time performance while maintaining reasonable accuracy for many applications.

The YOLO family has evolved through multiple iterations, with each version improving upon speed-accuracy trade-offs. YOLOv11 [26], in particular, offers several model variants ranging from nano (yolo11n) to extra-large (yolo11x) configurations. The nano variant is specifically designed for resource-constrained environments, featuring significantly reduced parameters and computational requirements while preserving essential detection capabilities.

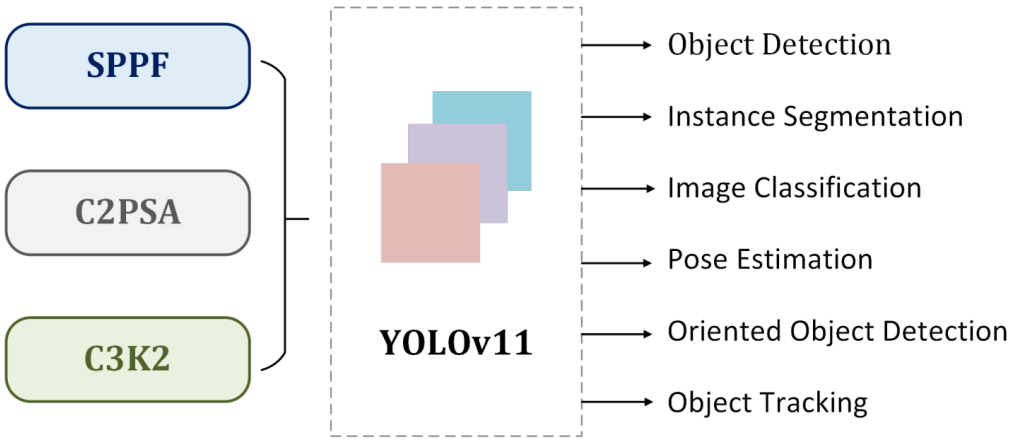

Figure 2.1: Key architectural modules in YOLO11 [26].

A significant advancement in YOLOv11 is the integration of thhe C2PSA (Convolutional block with Parallel Spatial Attention) component, which enhances spatial attention capabilities beyond previous YOLO iterations. The C2PSA block enables the model to focus more effectively on critical regions within images by inaplementing parallel spatial attention mechanisms. This enhancement is particularly beneficial for detecting objects of varying sizes and positions, addressing common challengees in complex visual environments with partially occluded or small objects. The retention of the Spatial Pyramid Pooling - Fast (SPPF) blockfrom previous versions, combined with the new C2PSA component, creates a compreehensive feature processing pipeline that balances computational efficiency with enhancced spatial awareness.

For edge-based human monitoring applications, YOLOv11n provides an optimal balance between detection performance and computational efficiency. Its lightweight architecture enables deployment on edge devices for real-time person detection, serving as the foundation for subsequent tracking and Re-ID processes in distributed

camera networks.

2.2.2 Object tracking

Person Re-ID systems face significant challenges when relying solely on frame-by-frame analysis. Individuals frequently lose their visual identity due to various factors including occlusions from other people or objects, rapid movement causing motion blur, and temporary disappearance from camera coverage areas. While deep learning-based feature extraction models can effectively capture contextual information and compute discriminative identity embeddings, frame-based matching approaches often suffer from identity fragmentation—where the same person receives multiple different identities across consecutive frames.

To address these limitations, tracking mechanisms play a crucial role in maintaining identity consistency over temporal sequences. Unlike existing methods that perform Re-ID independently for each frame, tracking-based approaches maintain continuous identity associations across time. This temporal continuity significantly outperforms computationally expensive alternatives such as query-driven region proposals [27] and graph-based retrieval methods [28], which become prohibitively costly in large-scale deployment scenarios. By leveraging tracking, our system can efficiently associate multiple detections of the same person, substantially reducing redundant identity searches while improving real-time processing capabilities.

The development of multi-object tracking (MOT) algorithms has evolved through several generations, each addressing specific limitations of previous approaches.

Multi-object tracking has evolved through several approaches, each with distinct trade-offs. SORT [29] provides efficient tracking by integrating object detection with motion prediction, but struggles with complex movement patterns and fast-paced scenarios. To address these limitations, DeepSORT [30] incorporates appearance features via a pre-trained Siamese network, improving performance in dense environments where motion alone is inadequate. However, this enhancement introduces dependency on embedding quality and computational complexity, making it susceptible to visual disturbances. FairMOT [31] advances this paradigm by merging detection and tracking into a unified architecture that simultaneously produces detection outputs and Re-ID features for enhanced multi-object tracking. This unified approach, while effective, demands significant computational resources and requires careful optimization between detection and Re-ID objectives, ultimately compromising processing speed. Alternative solutions include MMTracking [32], which provides a versatile framework supporting multiple advanced algorithms but necessitates substantial parameter optimization.

ByteTrack [33] represents a breakthrough in tracking methodology, delivering exceptional performance without requiring dedicated appearance models, thereby maintaining high processing speeds particularly in crowded environments. This approach achieves an optimal trade-off between real-time processing and tracking reliability. Consequently, our framework employs ByteTrack for maintaining pedestrian identity consistency across video frames, significantly enhancing ID assignment precision. This capability proves essential in dense scenarios where overlapping persons create significant challenges for camera-based identification systems.

2.2.3 Feature extraction

Feature extraction serves as the cornerstone of person Re-ID systems, transforming raw image data into compact descriptors suitable for robust identity matching. To enhance efficiency, especially for deployment on resource-constrained edge devices, recent research has focused on designing specialized lightweight feature extraction architectures.

For instance, Wang et al. introduced the Attention Knowledge-distilled Lightweight Network (ADLN) [34], specifically tailored for edge applications. ADLN utilizes a dimension interaction attention module to improve channel-wise feature representation, complemented by self-distillation that transfers learned attention patterns from deeper layers to shallower ones. Employing a combination of cross-entropy, weighted triplet, and center loss, ADLN effectively minimizes intra-class variability, achieving competitive accuracy on widely-used benchmarks such as Market-1501 and DukeMTMC-ReID, while significantly reducing computational complexity.

Building on similar objectives, Gao et al. presented a joint attention-based Re-ID model [35]. This model emphasizes both global and local pedestrian features through integrated attention modules, achieving precise and discriminative feature extraction that aligns well with realistic surveillance scenarios. This balance between accuracy and computational efficiency makes the model highly suitable for edge deployment.

Recent advances have also explored lightweight Transformer-based architectures, notably LightAMViT [36]. By streamlining self-attention mechanisms through K-means-based token clustering and adaptive weighted pooling, LightAMViT significantly reduces computational demands compared to traditional Vision Transformers. This approach provides a practical alternative, blending the strong representational capability of Transformers with the computational efficiency required by edge devices.

Complementing these specialized architectures, broader efforts in mobile-optimized methodologies such as MobileNetV2 [37] and MobileNetV3 [38] have gained prominence

in edge-based AI applications. These networks use depth-wise separable convolutions to significantly cut down computational overhead while preserving accuracy. Additionally, Squeeze-and-Excitation Networks (SE-Net) [39] introduce adaptive channel-wise recalibration mechanisms, further enhancing model efficiency by focusing computation on informative image regions.

However, despite these general improvements, generic lightweight architectures often lack specific optimizations required by the complexities of person Re-ID tasks. This gap has motivated dedicated research into architectures explicitly designed for Re-ID applications.

Among these specialized designs, OSNet (Omni-Scale Network) [8] represents a significant advancement. With only 2.2 million parameters, OSNet is substantially smaller than traditional methods like ResNet-50, which typically use around 24 million parameters, making it highly suitable for edge deployment. OSNet introduces innovative omni-scale feature learning through a novel building-block design, effectively capturing multi-scale information without the heavy computational cost typically associated with traditional multi-scale approaches. Its use of depth-wise separable convolutions and channel-shuffling operations maintains computational efficiency while preserving strong discriminative capability.

Expanding on OSNet's approach, LightMBN (Lightweight Multi-Branch Network) [9] further advances edge-friendly Re-ID through a multi-branch design tailored for constrained environments. LightMBN integrates efficient channel-wise and spatial attention mechanisms, adaptively emphasizing informative features without significantly increasing computational load. This enables robust identity matching with minimal resource usage.

2.2.4 Image classification

Traditionally, image classification and person Re-ID have been considered distinct tasks, with limited overlap. Person Re-ID focuses on matching individuals across different camera views, while image classification typically identifies broad object categories. However, when considering scenarios with a large search space—potentially thousands of identities—performing direct identity matching can be computationally expensive and slow. Under these conditions, reducing the search space using easily classifiable human metadata, such as gender, becomes beneficial by limiting the number of candidates considered, thus improving retrieval speed and accuracy.

Commonly used image classification models include well-established convolutional neural networks (CNNs) such as ResNet [40], known for its deep residual learning framework that significantly improves accuracy in classification tasks. MobileNet [41]

offers lightweight architectures optimized for resource-constrained devices, leveraging depth-wise separable convolutions to reduce computational costs. More recently, Vision Transformers (ViT) [42] have introduced transformer-based architectures to image classification, leveraging self-attention mechanisms to capture global relationships, achieving impressive accuracy at the cost of higher computational demands.

To effectively balance accuracy and computational efficiency in edge-based systems, EfficientNet [43] was introduced. EfficientNet leverages a compound scaling method to optimally adjust network depth, width, and resolution. Among its variants, EfficientNet-B0 stands out due to its particularly compact structure and excellent performance, making it ideal for deployment on lightweight edge devices. By employing EfficientNet-B0 for gender classification in our Re-ID pipeline, we effectively filter candidate matches by gender, significantly reducing computational overhead and improving the efficiency of subsequent identity matching stages.

2.2.5 Message queue

In hybrid edge-server deployments for person Re-ID systems, efficiently handling data flow between distributed devices is crucial. A reliable message queue system helps manage communication among edge devices, such as cameras and IoT sensors, and central processing servers. Message queues facilitate asynchronous and stable data transfer, allowing edge devices to process data locally without delays caused by direct server interactions.

Several widely-used message queuing solutions exist, such as Apache Kafka, RabbitMQ, and ZeroMQ, each with unique strengths:

RabbitMQ [44] is a well-established messaging broker that supports complex routing patterns and guaranteed message delivery. It is particularly effective for transactional systems where reliable delivery and message acknowledgment are vital. However, RabbitMQ can face scalability challenges when managing the high-throughput streaming data common in video-based person Re-ID scenarios.

ZeroMQ [45] is lightweight, extremely fast, and suitable for direct, point-to-point messaging. Its simplicity and speed make it ideal for real-time data transmission, but it lacks built-in persistence and fault tolerance mechanisms necessary for hybrid edge-server deployments that require robust data management over unreliable networks.

Considering these aspects, **Apache Kafka** [46] emerges as the most suitable choice for hybrid edge-server person Re-ID systems. Kafka excels in handling real-time streaming data, providing strong scalability, reliability, and fault-tolerant

message persistence. It efficiently processes high volumes of continuous data, such as features extracted from surveillance video streams, ensuring seamless integration between multiple distributed edge nodes and centralized processing servers.

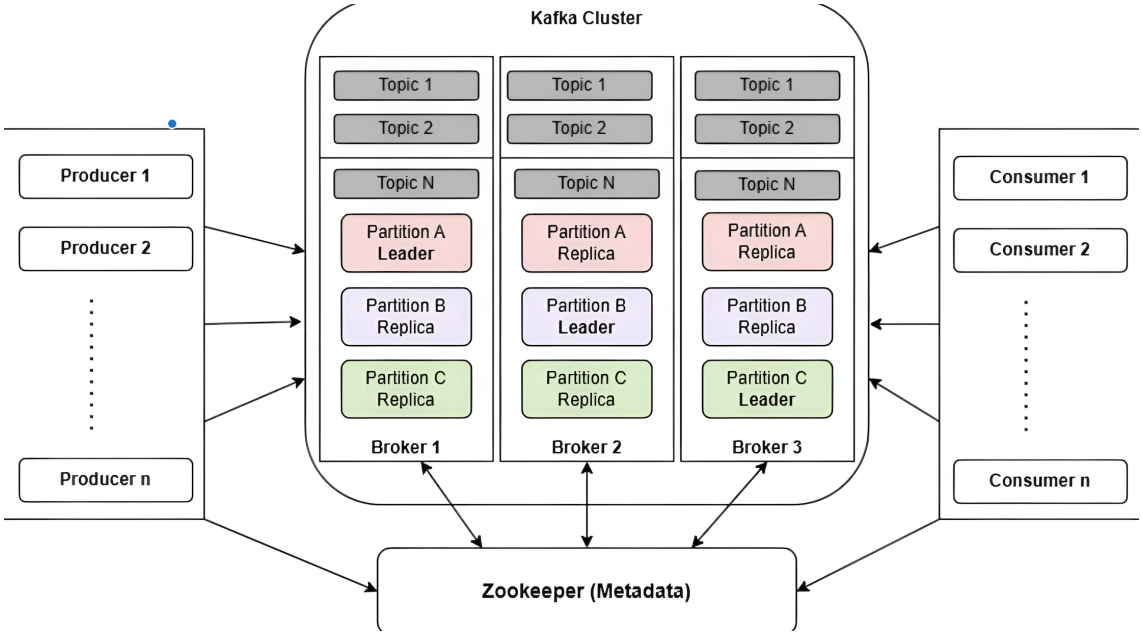

Figure 2.2: Kafka Architecture illustrating Producers, Consumers, Topics, Partitions, and Zookeeper [47].

In Kafka's architecture, *producers* generate data streams—such as encoded video frames in byte format, video metadata, and detected human bounding boxes—and send these streams to Kafka's storage system. In the context of person Re-ID, producers correspond to edge devices (e.g., surveillance cameras or edge processors). On the other hand, *consumers* are processes initiated on central servers, which simultaneously subscribe to these data streams to retrieve and perform further processing, such as feature extraction, identity matching, and analytics. Kafka organizes these data streams into logical channels known as *topics*, each representing a specific category or type of data. To enhance scalability, each topic is divided into multiple *partitions*, enabling parallel processing and improved fault tolerance across distributed environments.

This structured approach ensures Kafka can effectively manage data flow and scale seamlessly, making it particularly suited to complex, distributed Re-ID deployments.

2.2.6 Model serving framework

Deploying person Re-ID and lightweight classification models effectively in hybrid edge–server architectures requires selecting an appropriate model-serving framework. Several well-known solutions such as TorchServe, TensorFlow Serving, and NVIDIA Triton Inference Server have been widely adopted. TorchServe [48]

is popular due to its native support and ease of use with PyTorch models, featuring straightforward REST APIs and dynamic batching to enhance GPU throughput. However, it lacks the ability to concurrently serve multiple instances of the same model efficiently on a single GPU, limiting maximal GPU utilization for lightweight models. TensorFlow Serving [49], while efficient in serving TensorFlow models, presents additional complexity for PyTorch-based workflows due to the necessity of model conversion. NVIDIA Triton Inference Server [50], meanwhile, excels at maximizing GPU usage through concurrent model execution and dynamic batching, enabling high throughput and efficient GPU resource allocation. Nevertheless, Triton's complexity and configuration overhead can present challenges, particularly for development teams prioritizing rapid deployment and ease of integration.

To balance these trade-offs, Ray Serve [51] has emerged as an attractive framework, offering notable flexibility and ease of integration. Specifically designed for scalable deployments, Ray Serve supports serving multiple models simultaneously, dynamically composing inference pipelines, and efficiently managing replicas to maximize GPU utilization. Additionally, it seamlessly integrates with FastAPI, allowing developers to directly embed their inference endpoints within Python-based microservice architectures. This combination of flexible model orchestration, autoscaling capabilities, and developer-friendly integration makes Ray Serve particularly suited for deploying lightweight person Re-ID models like OSNet and LightMBN, as well as efficient image classification models, within hybrid edge–server setups.

2.2.7 Containerization

Containerization is a foundational technology in modern DevOps and microservices architectures, enabling applications and their dependencies to be packaged into lightweight, isolated units. This approach simplifies development workflows and significantly improves service portability. By encapsulating services within containers, they can be deployed, scaled, and updated independently, which facilitates more efficient version control and lifecycle management in distributed systems.

Among the various container technologies, Docker [52] is a widely used tool for building and running containers. Developers define application environments in a 'Dockerfile', which ensures consistent behavior across both staging and production environments.

Kubernetes complements Docker by orchestrating these containers at scale. It automates key operational tasks, including deployment, scaling, failover, and version rollouts across clusters. In the context of a hybrid edge-server person Re-ID pipeline, Docker is employed to package models, inference code, and APIs as distinct services.

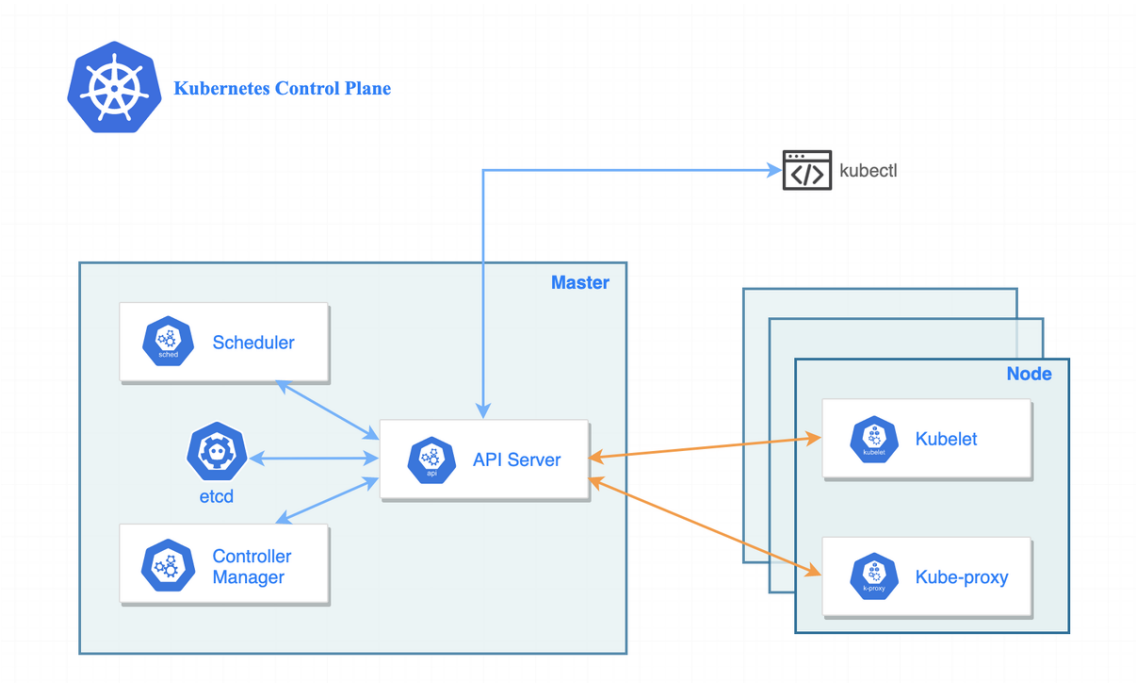

Figure 2.3: Kubernetes components [53].

Kubernetes then orchestrates these services—such as Kafka brokers, Ray Serve instances, and vector databases—to ensure system-wide reliability, scalability, and streamlined version control.

2.2.8 Vector Database

Traditional databases, such as OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) systems, are optimized for structured queries and aggregated data operations. However, these systems typically face limitations when handling high-dimensional, continuous data like image feature embeddings. In contrast, **vector databases** are specifically designed to efficiently store and query vector representations, enabling rapid similarity searches based on distance metrics, notably cosine similarity.

Although widely used in Natural Language Processing (NLP) tasks like Retrieval-Augmented Generation (RAG), vector databases also significantly contribute to the person Re-ID domain by facilitating efficient identity matching. In person Re-ID systems, each individual's identity is stored as a high-dimensional feature vector—commonly 512 or 1024 dimensions—depending on the chosen deep learning feature extractor. During the retrieval stage, the system queries this vector database with a new feature vector to identify the most similar stored identities.

Several popular vector databases have emerged, each with distinct characteristics:

- **Faiss**, developed by Meta, is a highly efficient C++ library offering both

exact and approximate nearest-neighbor search using indexing strategies like Hierarchical Navigable Small World (HNSW), product quantization, and GPU acceleration [54].

- **Milvus** is a scalable, distributed vector database built upon Faiss and hnswlib, supporting vast volumes of vectors and hybrid indexing methods optimized for distributed deployments [55].
- **ChromaDB** is lightweight and optimized for NLP embeddings, providing simplicity and ease of use, though it may not offer the highest performance for large-scale or complex metadata queries [56].
- **Redis**, traditionally a key-value store, now includes vector search capabilities, offering speed and convenience; however, it provides limited support for advanced metadata filtering and complex query patterns [57].
- **Qdrant**, a Rust-based vector database, excels in hybrid queries due to its built-in support for efficient metadata filtering combined directly with vector search. Its HNSW indexing structure integrates metadata filtering during the search, significantly reducing query latency compared to post-filtering approaches [58].

Vector database indexes, such as HNSW [59], differ fundamentally from traditional database indexes. Rather than focusing on exact matches or sorted queries, they organize vectors into graph-based structures optimized for approximate nearest-neighbor searches in high-dimensional spaces. The most commonly used distance metric in such scenarios is cosine similarity, which measures the angular similarity between vectors.

Among these alternatives, **Qdrant** aligns most closely with the needs of our hybrid edge-server person Re-ID pipeline. Its capability to perform rapid cosine similarity searches with integrated metadata filtering—such as gender classification obtained via classification model — enables significant reductions in search space, thereby enhancing retrieval speed, accuracy, and overall computational efficiency.

CHAPTER 3. METHODOLOGY

Building upon the theoretical foundations established in Chapter 2, this chapter presents a comprehensive examination of (i) the person Re-ID module architecture and its integration within the hybrid edge-server management system, (ii) detailed specifications of the proposed lightweight Re-ID module optimized for SME deployment, (iii) edge device hardware implementation strategies, (iv) the deployment of microservices frameworks and (v) how the system utilizes person metadata (gender) to enhance the efficiency of identity retrieval processes.

3.1 Overview

3.2 The proposed AI module

3.2.1 Human detection

a, Pre-processing

b, Detecting with YOLOv5n

c, Non-maximum suppression

3.2.2 Human feature extraction

a, Pre-processing

CHAPTER 4. EXPERIMENTAL RESULTS

CHAPTER 5. CONCLUSIONS AND FUTURE WORKS

REFERENCE

[1] S. PRO, *Hành trình nâng tầm trải nghiệm khách hàng ngành f&b*, 2024. **url**: https://soipro.vn/hanh-trinh-nang-tam-trai-nghiem-khach-hang-nganh-fb/.

[2] *Rising costs threaten vietnam's f&b profitability*. **urlseen** 2025. **url**: https://www.vietdata.vn/post/the-f-b-industry-is-seeing-its-profits-disappear-due-to-rising-costs, TheLEADER.

[3] V. News, *Nearly 60 per cent of food and beverage companies reported decline in revenue in 2024*, 2024. **url**: https://vietnamnews.vn/economy/1694177/nearly-60-per-cent-of-food-and-beverage-companies-reported-decline-in-revenue-in-2024.html.

[4] Y. Guo, W. Zhang, L. Jiao, S. Wang, S. Wang **and** F. Liu, "Visible-infrared person re-identification with region-based augmentation and cross modality attention," *Scientific Reports*, **jourvol** 15, **may** 2025. DOI: 10.1038/s41598-025-01979-z.

[5] W. Qin, Y. Li, J. Zhang, X. Wen, J. Guo **and** Q. Guo, "Attention-based hybrid contrastive learning for unsupervised person re-identification," *Scientific Reports*, **jourvol** 15, **april** 2025. DOI: 10.1038/s41598-025-97818-2.

[6] J. Yan, Y. Wang, X. Luo **and** Y.-W. Tai, *Fusionsegreid: Advancing person re-identification with multimodal retrieval and precise segmentation*, 2025. arXiv: 2503.21595 [cs.CV]. **url**: https://arxiv.org/abs/2503.21595.

[7] K. Niu **andothers**, *Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models*, 2025. arXiv: 2502.19958 [cs.CV]. **url**: https://arxiv.org/abs/2502.19958.

[8] K. Zhou, Y. Yang, A. Cavallaro **and** T. Xiang, *Omni-scale feature learning for person re-identification*, 2019. arXiv: 1905.00953 [cs.CV]. **url**: https://arxiv.org/abs/1905.00953.

[9] F. Herzog, X. Ji, T. Teepe, S. Hormann, J. Gilg **and** G. Rigoll, "Lightweight multi-branch network for person re-identification," **in***2021 IEEE International Conference on Image Processing (ICIP)* IEEE, **september** 2021, 1129–1133. DOI: 10.1109/icip42928.2021.9506733. **url**: http://dx.doi.org/10.1109/ICIP42928.2021.9506733.

[10] G. V. Research, *Edge ai market size, share & growth | industry report, 2030*, 2024. **url**: https://www.grandviewresearch.com/industry-analysis/edge-ai-market-report.

[11] "2024 tech trends: How to reduce friction in the retail experience," *BizTech Magazine*, 2024. **url**: https://biztechmagazine.com/article/2024/03/2024-tech-trends-how-reduce-friction-retail-experience.

[12] "Person re-identification network based on edge-enhanced feature extraction and inter-part relationship modeling," *Applied Sciences*, **journvol** 14, **number** 18, **page** 8244, 2024. **url**: https://www.mdpi.com/2076-3417/14/18/8244.

[13] Viso.ai, *Edge intelligence: Edge computing and ml (2025 guide)*, 2024. **url**: https://viso.ai/edge-ai/edge-intelligence-deep-learning-with-edge-computing/.

[14] Wikipedia, *Microservices*, 2024. **url**: https://en.wikipedia.org/wiki/Microservices.

[15] "Person re-identification microservice over artificial intelligence internet of things edge computing gateway," *Electronics*, **journvol** 10, **number** 18, **page** 2264, 2021. **url**: https://www.mdpi.com/2079-9292/10/18/2264.

[16] "Distributed implementation for person re-identification," **in***IEEE Conference Publication* IEEE, 2015. **url**: https://ieeexplore.ieee.org/document/7288501/.

[17] "Mded-framework: A distributed microservice deep-learning framework for object detection in edge computing," *Sensors*, **journvol** 23, **number** 10, **page** 4712, 2023. **url**: https://www.mdpi.com/1424-8220/23/10/4712.

[18] "Video-based person re-identification based on distributed cloud computing," *Journal of Artificial Intelligence and Technology*, 2022. **url**: https://ojs.istp-press.com/jait/article/view/13.

[19] Splunk, *What are distributed systems?* 2024. **url**: https://www.splunk.com/en_us/blog/learn/distributed-systems.html.

[20] R. Girshick, J. Donahue, T. Darrell **and** J. Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, 2014. arXiv: 1311.2524 [cs.CV]. **url**: https://arxiv.org/abs/1311.2524.

[21] K. O'Shea **and** R. Nash, *An introduction to convolutional neural networks*, 2015. arXiv: 1511.08458 [cs.NE]. **url**: https://arxiv.org/abs/1511.08458.

[22] R. Girshick, *Fast r-cnn*, 2015. arXiv: 1504.08083 [cs.CV]. **url**: https://arxiv.org/abs/1504.08083.

[23] S. Ren, K. He, R. Girshick **and** J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, 2016. arXiv: 1506.01497 [cs.CV]. **url**: https://arxiv.org/abs/1506.01497.

[24] G. Merz **andothers**, "Detection, instance segmentation, and classification for astronomical surveys with deep learning implementation and demonstration with hyper suprime-cam data," *Monthly Notices of the Royal Astronomical Society*, **jourvol** 526, **number** 1, 1122–1137, **september** 2023, ISSN: 1365-2966. DOI: 10.1093/mnras/stad2785. **url**: http://dx.doi.org/10.1093/mnras/stad2785.

[25] M. Tan, R. Pang **and** Q. V. Le, *Efficientdet: Scalable and efficient object detection*, 2020. arXiv: 1911.09070 [cs.CV]. **url**: https://arxiv.org/abs/1911.09070.

[26] R. Khanam **and** M. Hussain, *Yolov11: An overview of the key architectural enhancements*, 2024. arXiv: 2410.17725 [cs.CV]. **url**: https://arxiv.org/abs/2410.17725.

[27] B. Munjal, S. Amin, F. Tombari **and** F. Galasso, *Query-guided end-to-end person search*, 2019. arXiv: 1905.01203 [cs.CV]. **url**: https://arxiv.org/abs/1905.01203.

[28] Z. Zhu, T. Huang, K. Wang, J. Ye, X. Chen **and** S. Luo, *Graph-based approaches and functionalities in retrieval-augmented generation: A comprehensive survey*, 2025. arXiv: 2504.10499 [cs.IR]. **url**: https://arxiv.org/abs/2504.10499.

[29] A. Bewley, Z. Ge, L. Ott, F. Ramos **and** B. Upcroft, "Simple online and realtime tracking," **in***2016 IEEE International Conference on Image Processing (ICIP)* IEEE, **september** 2016. DOI: 10.1109/icip.2016.7533003. **url**: http://dx.doi.org/10.1109/ICIP.2016.7533003.

[30] N. Wojke, A. Bewley **and** D. Paulus, *Simple online and realtime tracking with a deep association metric*, 2017. arXiv: 1703.07402 [cs.CV]. **url**: https://arxiv.org/abs/1703.07402.

[31] Y. Zhang, C. Wang, X. Wang, W. Zeng **and** W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, **jourvol** 129, **number** 11, 3069–3087, **september** 2021, ISSN: 1573-1405. DOI: 10.1007/s11263-021-01513-4. **url**: http://dx.doi.org/10.1007/s11263-021-01513-4.

[32] C. Lin, C. Yu, X. Xu **and** R. Wang, *Mmtracking: Trajectory tracking for uplink mmwave devices with multi-path doppler difference of arrival*, 2025. arXiv: 2503.16909 [eess.SP]. **url**: https://arxiv.org/abs/2503.16909.

[33] Y. Zhang **andothers**, *Bytetrack: Multi-object tracking by associating every detection box*, 2022. arXiv: 2110.06864 [cs.CV]. **url**: https://arxiv.org/abs/2110.06864.

[34] W. Jin, D. Yanbin **and** C. Haiming, "Lightweight person re-identification for edge computing," *IEEE Access*, **jourvol** 12, **pages** 75 899–75 906, 2024. DOI: 10.1109/ACCESS.2024.3405169.

[35] S. Jiao, J. Wang, G. Hu, Z. Pan, L. Du **and** J. Zhang, "Joint attention mechanism for person re-identification," *IEEE Access*, **jourvol** PP, **pages** 1–1, **july** 2019. DOI: 10.1109/ACCESS.2019.2927170.

[36] H. Dong, I. Kotenko **and** S. Dong, "A lightweight vision transformer with weighted global average pooling: Implications for iomt applications," *Complex & Intelligent Systems*, **jourvol** 11, **march** 2025. DOI: 10.1007/s40747-025-01842-8.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov **and** L.-C. Chen, *Mobilenetv2: Inverted residuals and linear bottlenecks*, 2019. arXiv: 1801.04381 [cs.CV]. **url**: https://arxiv.org/abs/1801.04381.

[38] A. Howard **andothers**, *Searching for mobilenetv3*, 2019. arXiv: 1905.02244 [cs.CV]. **url**: https://arxiv.org/abs/1905.02244.

[39] J. Hu, L. Shen, S. Albanie, G. Sun **and** E. Wu, *Squeeze-and-excitation networks*, 2019. arXiv: 1709.01507 [cs.CV]. **url**: https://arxiv.org/abs/1709.01507.

[40] K. He, X. Zhang, S. Ren **and** J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. **url**: https://arxiv.org/abs/1512.03385.

[41] A. G. Howard **andothers**, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, 2017. arXiv: 1704.04861 [cs.CV]. **url**: https://arxiv.org/abs/1704.04861.

[42] A. Dosovitskiy **andothers**, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. **url**: https://arxiv.org/abs/2010.11929.

[43] M. Tan **and** Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. **url**: https://arxiv.org/abs/1905.11946.

[44] RabbitMQ, *RabbitMQ Documentation*, Accessed: 2024-06-10, 2024. **url**: https://www.rabbitmq.com/documentation.html.

[45] ZeroMQ, *ZeroMQ Guide*, Accessed: 2024-06-10, 2024. **url**: https://zeromq.org/get-started/.

[46] Apache Kafka, *Apache Kafka Documentation*, Accessed: 2024-06-10, 2024. **url**: https://kafka.apache.org/documentation/.

[47] 200lab, *Kafka là gì?* 2024. **url**: https://200lab.io/blog/kafka-la-gi.

[48] P. Team, *TorchServe Documentation*, Accessed: 2025-06-10, 2025. **url**: https://pytorch.org/serve/.

[49] T. Team, *TensorFlow Serving Documentation*, Accessed: 2025-06-10, 2025. **url**: https://www.tensorflow.org/tfx/guide/serving.

[50] NVIDIA, *NVIDIA Triton Inference Server Documentation*, Accessed: 2025-06-10, 2025. **url**: https://developer.nvidia.com/nvidia-triton-inference-server.

[51] Ray Team, *Ray Serve Documentation*, Accessed: 2025-06-10, 2025. **url**: https://docs.ray.io/en/latest/serve/index.html.

[52] Docker, *Docker Documentation*, 2024. **url**: https://docs.docker.com/.

[53] Nishan Baral, *Kubernetes Architecture Explained in Layman's Terms*, 2023. **url**: https://www.linkedin.com/pulse/kubernetes-architecture-explained-laymans-terms-nishan-baral-ydjyc.

[54] Meta Research, *Faiss Documentation*, 2024. **url**: https://github.com/facebookresearch/faiss.

[55] Zilliz, *Milvus Documentation*, 2024. **url**: https://milvus.io/.

[56] Chroma, *ChromaDB Documentation*, 2024. **url**: https://www.trychroma.com/.

[57] Redis, *Redis Vector Similarity Search*, 2024. **url**: https://redis.io/docs/latest/develop/get-started/vector-database/.

[58] Qdrant, *Qdrant Documentation*, 2024. **url**: https://qdrant.tech/documentation/.

[59] Pinecone, *Hierarchical Navigable Small Worlds (HNSW)*, 2024. **url**: https://www.pinecone.io/learn/series/faiss/hnsw/.