Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 258 (2025) 455–466

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

A Comprehensive Deep Learning Model for Improved Person Re-identification Using Multi-Camera Streaming Pipeline

Patel Devarshi[a] , Chauhan Yash[b], Meghna B. Patel[c,]*, Dhaval K. Raval[d]

[a,b,d] *Department of Computer Science, Ganpat University, Kherva, Gujarat, India, 384012*
[c] *Acharya Motibhai Patel Institute of Computer Studies, Ganpat University, Kherva, Gujarat, India, 384012*

Abstract

The growing importance of multi-camera person Re-ID reflects its significance in computer vision. While single-camera tracking has advanced, maintaining identification across multiple cameras remains challenging due to factors like occlusion, appearance changes, camera motion, and lighting conditions. In this proposed research work, present a comprehensive multi-camera person Re-ID process designed to address these challenges by leveraging advanced detection and tracking methodologies. The process begins with the utilization of the state-of-the-art YOLOv8n and lightweight object detection model, specifically for person detection. This model efficiently identifies and localizes individuals within the video frames captured by multiple cameras, each covering a unique perspective within the monitored environment. The system integrates DeepSORT trained on the MARS dataset for multiple object tracking (MOT) to maintain consistent identification across video frames. Following person detection and tracking, a custom transformer model built on the TorchReID framework is employed to analyze and compare bounding boxes of detected persons, ensuring that the identified individuals are accurately associated with existing IDs. The Re-ID process is further enhanced by custom trained OSNet_x1_0 and ResNet50 models. These models, trained on a combination of the Market1501 dataset and the UNI09 custom dataset, extract rich and discriminative feature embeddings. The integration of these models has achieved high accuracy in identifying individuals across multiple camera views. Notably, OSNet_x1_0 achieves 98.4% mAP and ResNet50 reaches 96.1% mAP on Market1501. The integration of these models into the Re-ID pipeline has achieved high accuracy in identifying individuals across multiple camera views. Experimental results demonstrate the system's ability to maintain accurate person identification across cameras, achieving high consistency. This outcome highlights the effectiveness of proposed approach.

* Corresponding author. Tel.: +91 88499 71703;
E-mail address: meghna.patel@ganpatuniversity.ac.in

10.1016/j.procs.2025.04.281

Keywords: Person Re-identification; Computer Vision; MOT; Tracking; Surveillance;

Nomenclature

Re-ID	Re-Identification
OSNet	Omni-Scale Network
MOT	Multi Object Tracking
ResNet	Residual Network
GPU	Graphical Processing Unit
mAP	Mean Average Precision
CMC	Cumulative Match Characteristics
YOLO	You Look Only Once
TF	Transfer Learning
NMS	Non-Maximum Suppression

1. Introduction

Multi-camera person Re-ID is a critical task in the field of computer vision, with significant implications for surveillance, security, and various real-world applications. Person Re-ID aims to recognize and track a specific individual, maintain a consistent identity for a person across multiple sightings even when their appearance may vary due to different camera angles, lighting conditions, or the person's pose. It is crucial for several applications, particularly in surveillance and security: It allows for tracking individuals across large areas with multiple cameras, provide a view of their movements, Helps in locating persons of interest in crowded areas, Major use case in finding location of missing person. The model will automate the process of person detection and tracking between frames, reducing the need for manual intervention [1].

The ReID system is capable of reconstructing the order of events in various places when it receives video data from several cameras. Investigators can follow a suspect's activities across time thanks to this feature, which gives them a complete picture of what they did before, during, and after the incident. Finding an individual of interest can occasionally be challenging due to significant shifts in body language and occlusion [2]. As surveillance networks expand, the system must scale to manage a large number of cameras. This scaling introduces significant computational challenges, as the system must process and analyze extensive data streams from multiple sources. There are several other factors that also matter except ReID, bounding box tracker used to maintain the ID of trajectory present in frame. Ensure that the tracker receives accurate and timely input from the detection stage to establish a solid foundation for tracking. Maintain the trajectory of individuals across frames by predicting their future positions based on their current trajectory [3]. MOT after detection is crucial for person ReID as it ensures that IDs are consistently maintained for individuals across frames. Without effective MOT, the model will struggle a lot to track individuals accurately, leading to errors in ID assignment. This algorithm is designed to follow each person's trajectory across frames and manage their identities [4], it is done in two steps object detection and target association. There are few options including ByteTrack, BotTrack, SORT and DeepSORT, which integrates appearance features for improved accuracy and long-term tracking.

The input image or video’s detected part is resized to fixed size to match input standards required by model. The input then passed through a CNN backbone. The backbone network serves to extract key features from the provided image input. That was basic workflow of object detection, each model have their own working methodology. Each bounding box prediction is accompanied by class probabilities. The model predicts the probability of each class being present in the bounding box, allowing it to classify the detected objects. It has common problem of overlapping and duplicate bounding box, to solve that issue there is term called NMS, which retain only bounding boxes that have highest confidence of class label. There already exist a few review articles on person detection [5], [6]. Each of them use different detection models and methods. The paper [5] used YOLO object detection. For Instance, in [6] the MobileNet and SSD detector are used for object detection. The structure of the paper is organized as follows. Section

2 reviews various methods employed to achieve ReID, person detection and types of MOT available in the ReID area. In Section 3, it provides the materials and methods utilized to achieve multi-camera person ReID. In Section 4, the methodology of paper is discussed in detail, In Section 5 it shows the results. At Section 6 the conclusion is given.

2. Related Works

2.1. Person Detection

The initial process of our work involves capturing footage from multiple camera angles. This gathered footage is then fed into an object detection model for person detection. In the paper [7] , the authors discuss two types of object detection models. Single-stage detectors such as YOLO, SSD, M2Det and EfficientDet, along with two-stage detectors like Region-based CNN, Faster R-CNN, Mask R-CNN, and Cascade R-CNN, are widely used in object detection tasks. Single stage models are lightweight and faster in performance, while two-stage models are heavier and slower but offer higher accuracy compared to single-stage models.

The authors in [8] propose real-time person detection for edge devices, such as NVIDIA Jetson and Raspberry Pi, using the SSD MobileNetV2 model. Their approach combines feature extraction, deformation handling, occlusion management, and classification to achieve accurate pedestrian detection without sacrificing real-time performance. The work described in [9] introduces an automatic human detection system that utilizes the VGG16 [10] architecture for deep learning-based image processing. With an accuracy of 60%, this system effectively identifies humans in images. The authors of [11] propose the A system for automated person detection and tracking powered by deep learning techniques system. This system leverages the EfficientDet model for detecting and tracking people in frames. According to their claims, DLD-APDT achieves higher accuracy than Faster R-CNN and YOLOv3 [12] in terms of recall and precision.

2.2 ReID

Multi-camera multi-person tracking and ReID are crucial tasks in various scenarios, including indoor environments like operating rooms and retail stores. Some approaches have focused on integrating tracking and ReID tasks using trajectory-based methods [13] and appearance embeddings clustering [14], These techniques aim to overcome challenges such as occlusions and misidentification in crowded spaces. Previously, research explored the use of Implicit Shape Models and SIFT features for person ReID [15] as well as graph-based signature generation considering both color and spatial features [16], These approaches have shown promising results in various multi-camera scenarios, with some achieving up to 85.44% accuracy in ReID tasks [13] . In [17] a method is proposed for rapid deployment in new environments using labeled data from a single camera. Both labeled and unlabeled data are stored in memory, with the labels offering supervision. The unlabeled images are grouped into clusters, with the cluster labels also contributing to supervision. The model utilizes ResNet50, pre-trained on ImageNet, as its backbone for feature extraction. Distances between the generated features are calculated, and DBSCAN clustering is applied to create pseudo labels. This approach forms a unified framework, as it leverages labeled data from a single camera, despite the lack of multi-camera annotations for each identity. Authors in [18] explore a method for person re-identification across visible and thermal camera images using a deep residual CNN with a single input stream. Their approach involves capturing images from both visible and thermal cameras. To capture and correlate features from these diverse image types, they employ feature extraction techniques such as IPVT-1, IPVT-2, and IIPVT. These features are then processed through a CNN to identify common attributes. The paper [19] introduces a new training setting called Single-Camera Training (SCT), where each person appears in only one camera. SCT simplifies data collection since it doesn't require cross-camera annotations, unlike Fully-Supervised Training (FST). It also provides reliable labels compared to Unsupervised Training, which has no labels. This paper [20] introduces BUPTCampus, a large video-based visible-infrared person ReID that addresses limitations in existing datasets, The authors propose a two-stream framework with Generative Adversarial Networks to reduce modality gaps, a curriculum learning strategy to utilize the auxiliary set, and a temporal k-reciprocal re-ranking method for refining results. The Multi-Memory Matching (MMM) [21] framework enhances unsupervised based infrared person ReID by improving pseudo labels generation and cross-modality correspondence. It includes a Cross-Modality Clustering module, a Multi-Memory Learning and Matching (MMLM) module and Soft Cluster-level Alignment loss. The [22] uses OSNet to extract features, which helps achieve a good balance between better performance and model efficiency. The Position Attention Module

(PAM) focuses on specific details, while the Efficient Channel Attention (ECA) handles interactions between different channels. This approach reduces the number of parameters needed without losing accuracy. The authors [23] explore how combining vision transformers with traditional CNN models can improve person re-identification. They test various backbone models on well-known benchmark datasets. By integrating transformers with Residual, Dense, and PCB networks. The MICRO-TRACK [24] implemented entire person re-identification pipeline on NVIDIA Jetson AGX Xavier edge computing device and achieved real-time faster FPS.

2.3 MOT

Person ReID in Multi-Camera system involve matching individuals across different camera views for enhanced surveillance and security [25] .Since current trackers such ByteTrack [26] which excels in handling occlusions and high-density scenarios, BotTrack [27] which extends traditional methods for more complex environments, and FairMOT [28] mostly rely on higher processing power and use relatively simple data association methods. Evaluating the performance of tracker the MOTA (Mult-Object-Tracking-Accuracy) is utilized. The DeepSORT [29] use both appearance and motion information for data association could provide more accurate tracking and it has MOTA 61.4, which is better than previous version of DeepSORT called SORT [30] both have same final matching stage where they run IoU association on group of not confirm and unmatched tracks, especially in challenging scenarios. It uses model that is specifically trained for Video-based Person ReID called MARS [31], it extracts features from trajectory bounding boxes tracked to predict upcoming frame to maintain. MARS is video based person Re-ID dataset and it is an extension to Market1501 [32].

3. Materials and Methods

3.1 Multi-Camera ReID Architectures

Entire re-identification process is done under TorchReID [33], it is an open-source person re-identification library based on PyTorch, the main goal is to track people across in multi-camera. It offers wide range of dataset (Market1501, CUHK03, DukeMTMC, MSMT17 etc.), ImageNet pre-trained models architectures (ResNet, DenseNet, Inception-V4, Xception, IBN-NET, OSNet etc.), it provide platform for training and building custom transformer and engines. The ImageNet contain 14,197,122 annotated pictures based on the WordNet hierarchy. It is primarily used to supply a wide range of common objects for training and evaluating models. ImageNet transfer learning helps accelerate model training, improve performance, and reduce data requirements [34] [35].

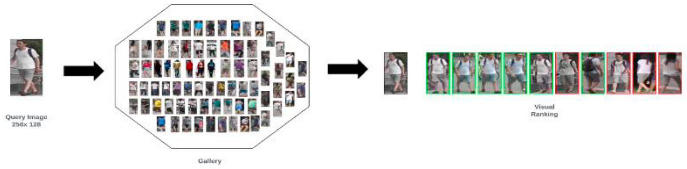

Fig 1. The process of visual ranking generation from a given query.

Person ReID generally operate using datasets that include gallery, query, and training data. The model will be trained and evaluated using datasets. The training data is required to train the model, while the query data is evaluated to evaluate the model's correctness.

The gallery contains images from both the training and query datasets, along with distracting or fake images to challenge the model's predictive capabilities. Upon receiving a query, the model searches the gallery for the top matches and ranks them according to their relevance to the query image. Visual ranking shown in ***Fig 1*** will give output of similar matchings of query, where red border box shows the false predictions and the green outlined box highlights the correct predictions based on similarity.

3.1.1 ResNet50 (Residual Network):-

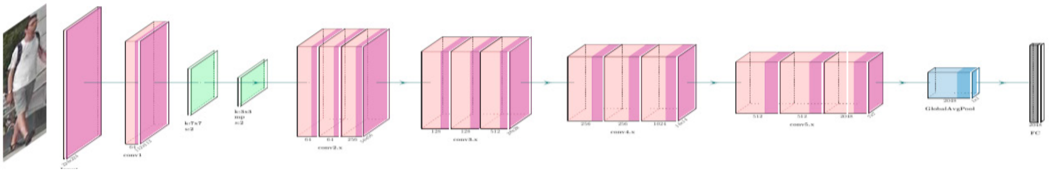

Fig 2. The architecture of ResNet50 Image Processing Model.

The ResNet-50 architecture is a deep convolutional neural network that addresses vanishing gradients, allowing for effective training of very deep networks using residual blocks. It receives a 256x128 pixel image as input, which passes through a series of layers to extract essential features as shown in ***Fig 2***. The first convolutional layer has a 7x7 kernel, a stride of 2, and 64 output channels, using a large receptive field to capture broad characteristics from the input image. This is followed by a max-pooling layer with a 3x3 kernel and a stride of 2, reducing the spatial dimensions of the feature maps. The feature maps then enter the first set of residual blocks, Conv2, which uses 1x1, 3x3, and 1x1 convolutions to adjust feature map counts, extract features, and restore map dimensions. The residual connection in these blocks enables the original input to bypass convolutions, ensuring better gradient flow. Subsequent stages Conv3, Conv4, and Conv5 follow a similar structure, with each increasing in depth and complexity. Conv3 outputs 128 feature maps, Conv4 outputs 256, and Conv5 produces 512, capturing increasingly complex features. The network then applies global average pooling to reduce spatial dimensions, resulting in a 2048-dimensional feature vector. This vector is processed by a fully connected layer that maps it to the final output classes.

3.1.2 OSNet_x1_0 (Omni-Scale Network):-

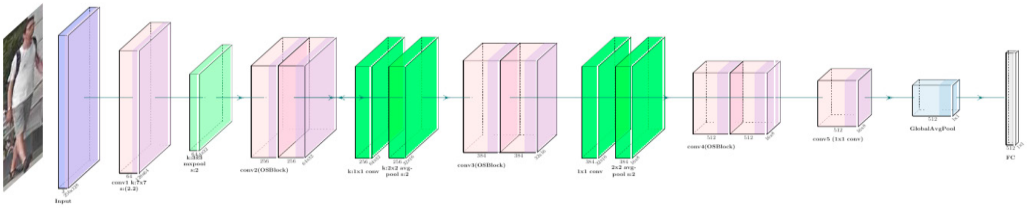

Fig 3. Architecture of Omni-Scale Network.

The OSNet_x1_0 architecture, designed for person re-identification, combines efficiency and effectiveness with convolutional, pooling, and transition layers arranged logically to extract features from input images as shown in ***Fig 3***. The input layer takes a 256x128x3 image and passes it to the first convolutional layer, conv1, with a 7x7 kernel and a stride of (2, 2), producing 64 output channels. This conv1 layer detects various features in the input, while the kernel and stride sizes reduce spatial dimensions. Following conv1, the first transition layer applies a max pooling operation with a 2x2 kernel and a stride of 2, further reducing spatial dimensions and capturing significant features from previous layers. The processed feature map, consisting of an OSBlock with 256 output channels, is obtained from the last layer and sent into the conv2 input. OSBlocks are specialized convolutional blocks created by [36] [37] to facilitate effective feature representation and learning shown in

***Fig* 4**. To put it simply, it uses a variety of convolutional techniques with various kernel sizes and skip connections to gather a large amount of feature data. It receives input, processes it using a 1x1 kernel, and then feeds the result into a lite 3x3 layer, where lite 3x3 is used for making neural networks lightweight. Here is an explanation. To learn residual mapping $\tilde{x}$ such that the output y is the sum of the input x and residual $\tilde{x}$.

$$y = x + \overset{o}{x} \; where \; \overset{o}{x} = f(x) \qquad (1)$$

Where f represent the Lite 3x3 layer that responsible for learning single-scale features on the scale of 3.

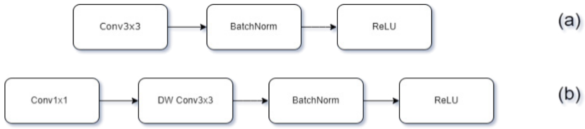

Fig 4. (a) Standard 3x3 convolution and (b) lite 3x3 convolution.

Finally, OSNet_x1_0 processes the image with a 1x1 convolution layer, followed by ReLU to produce the output. The initial 1x1 layers reduce feature dimensions, while the final ones restore them. After conv2, a second transition layer applies 1x1 convolution with a 2x2 average pooling, minimizing feature maps and reducing spatial resolution. The output then goes to conv3 (an OSBlock) with 384 channels, extracting complex features. A third transition layer follows conv3, similar to previous transitions. The next layer, conv4, is another OSBlock with 512 output channels, capturing rare features. Conv5, a 1x1 layer with 512 channels, refines features from conv4 and prepares them for global average pooling, which reduces spatial dimensions to 1x1, yielding a 512-dimensional feature vector that represents the entire image. The final fully connected layer takes this vector and maps it to the desired output classes, completing the classification task.

3.2 YOLOv8 with DeepSORT:

YOLOv8 [38] [39] is an optimized version of the YOLO family [40] of object detection models. It builds on the success of its predecessors by introducing several enhancements that improve both speed and accuracy. The architecture is separated into three main parts: the backbone, the neck, and the head. The backbone is used for extracting essential features from the input image. It serves as the model’s feature extractor, processing the image to produce a collection of feature maps that highlight crucial aspects such as edges, textures, and object outlines. The backbone typically consists of multiple convolutional layers that progressively reduce the spatial dimensions of the input while increasing the depth (number of channels) to capture more abstract, high-level features. YOLOv8 uses a modified CSPDarknet53 [41] architecture as its backbone. The neck refines and combines features from the backbone, creating multi-scale feature maps to detect objects of varying sizes. It includes feature fusion and upsampling to merge information from different backbone levels, capturing both fine details and contextual information. In YOLOv8, the C2f module replaces the CSPLayer from YOLOv5. The head performs the final detection, processing features from the neck to predict bounding boxes, classes, and objectness scores. It uses convolutional layers to output bounding box coordinates, class probabilities, and objectness scores, indicating the likelihood of an object in each box.

DeepSORT expands the SORT algorithm by including appearance information to improve object tracking across frames, particularly effective in person ReID and MOT. Trained on the MARS dataset, it uses YOLOv8 for person detection in each video frame. The Kalman filter predicts each object's next position based on its trajectory, while the appearance descriptor generates embeddings for detected objects. The algorithm matches new detections to existing tracks using motion and appearance information, and the Kalman filter updates each track state with the matched detection.

4. Proposed Methodology

In a multi-camera surveillance system, video feeds from multiple points within a monitored area are merged into a single video stream, providing centralized processing and analysis from numerous views. Each camera records real-time footage, which is crucial to identifying people. The merged video stream goes through to person detection using the YOLOv8n model, which recognizes and detects humans inside video frames by creating bounding boxes around their places.

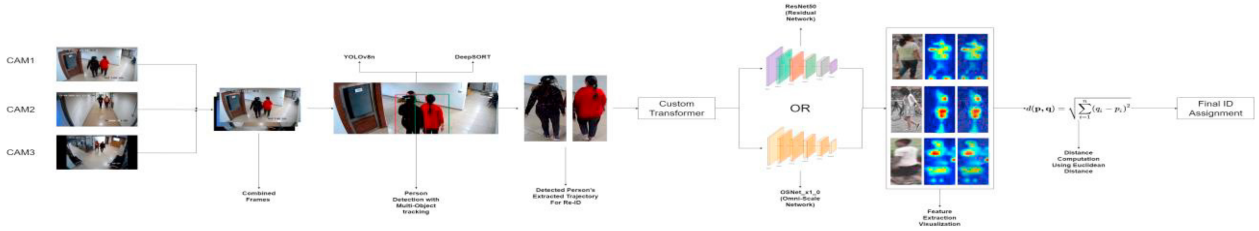
CAM1
CAM2
CAM3
Combined Frames
YOLOv8n
DeepSORT
Person Detection with Multi-Object tracking
Detected Person's Extracted Trajectory For Re-ID
Custom Transformer
ResNet50 (Residual Network)
OR
OSNet_x1_0 (Omni-Scale Network)
Feature Extraction Visualization
$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$
Distance Computation Using Euclidean Distance
Final ID Assignment

Fig 5. The flow of proposed methodology for Multi-Camera person re-identification using object detection model and tracking.

To ensure that each identified individual is accurately recognized over multiple frames inside the same camera, the DeepSORT algorithm is used for tracking. This method creates and retains unique identities (IDs) for each observed person, allowing for reliable monitoring of individuals across time. DeepSORT's tracking data is used to extract individual trajectories, exposing movement patterns across video frames. To identify individuals across several cameras, a custom transformer is utilized in combination with models such as OSNet_x1_0 or ResNet50 to extract robust feature embeddings for each recognized individual. These feature embeddings represent important visual features including look, clothes, and posture. The system evaluates individual similarity by comparing the feature vectors of observed persons using the distance metric. The Euclidean distance between two vectors of features is determined using the below.

$$d = \sqrt{\sum_{i=1}^{n} (qf_i - gf_i)^2} \tag{2}$$

Here, qf and gf represent the feature vectors for the query and gallery images, n is the number of feature in each vectors. Each person is granted a unique ID based on the computed distances, ensuring that the same individual keeps the same ID across multiple cameras and time frames. The model uses Softmax loss function for classifying individuals during training. It outputs a probability distribution over classes, and the model learns to classify input features correctly during training. But in inference it focuses on the embeddings rather than classification labels.

$$loss = -\sum_{i=1}^{n} y_i \log(p_i) \tag{3}$$

y_i represent the ground truth label, and p_i is the probability assigned by the model for that label. This method allows for accurate person ReID across several cameras in a complex monitored area, hence increasing the overall efficacy of the surveillance system.

5. Results and Discussions

In this research, an experiment was carried out using the Market-1501, MARS and UNI09 (custom) dataset for person ReID. A transfer learning-based OSNet and ResNet50 model was proposed, pre-trained on ImageNet and fine-tuned, surpassing supervised domain adaptation models. The next sections go into the dataset attributes and results of the proposed technique, which uses Market1501, MARS, and UNI09 as shown in ***Table 1***. For the MARS dataset, we employed a transfer learning strategy with a pre-trained model that was available.

Table 1. Dataset Characteristics

Dataset	Market1501	UNI09	MARS
Persons	1501	9	1261
Cameras	6	5	6
Query Images	3368	23	-
Total Images	32668	7561	1067516

The evaluation of the proposed person Re-ID models in this study is based on standard evaluation metrics commonly used in the field, including the CMC curve, which measures the ranking accuracy at various ranks such as Rank-1 and Rank-5, and the mAP, which evaluates the overall retrieval performance. CMC Curve represents the chances of an ideal match for a given query will appear in the top ranks of the results. Rank-1 accuracy refers to the possibility that the right match is the most highly ranked result, while Rank-5 accuracy measures the probability that the correct match appears within the top five ranked results. mAP considers both the precision and recall across all ranked results. A higher mAP value indicates that the model is more effective in ranking. Our models utilize transfer learning, where they are initially pre-trained on the ImageNet dataset and then custom trained on the Market1501 and UNI09 datasets.

Table 2. Comparing the models results with existing methods and models on Market1501

Methods & Models	mAP	Rank-1	Rank-5
MSCAN [42]	57.5	80.3	-
DRNTLPReID [43]	68.4	96.0	-
GLAD [44]	73.9	89.9	-
PCB [45]	77.4	91.3	97.1
PyrNet [46]	81.7	93.6	-
ResNet50 + CircleLoss [47]	84.9	94.2	-
PCL-CLIP [48]	91.4	95.9	98.5
Supervised(ResNet50+MGN) [49]	91.9	96.9	-
DG-Net(RK) [50]	92.49	95.4	-
LDS (ResNet50+RK) [51]	94.89	96.17	-
DAAF-BoT [52]	95.0	96.4	-
ResNet50+RK [53]	95.3	96.4	-
BPBreID (RK) [54]	95.3	96.4	-
LightMBN (RR) [55]	95.3	96.8	-
Viewpoint-Aware Loss (RK) [56]	95.43	96.79	98.31
st-ReID (RE,RK) [57]	95.5	98.0	98.9
SOLIDER (RK) [58]	95.6	96.7	-
RGT&RGPR (RK) [59]	95.6	96.9	-
ResNet101+RK [60]	96.21	-	-
ResNet50+TF (ours)	**96.1**	**99.1**	**99.8**
OSNet_x1_0+TF (ours)	**98.4**	**99.6**	**99.8**

Table 2 presents a comparison of our models' performance against various existing methods and models on the Market1501 dataset. The OSNet_x1_0 model achieved a mAP of 98.4%, along with Rank-1 and Rank-5 accuracies of 99.6% and 99.8%, respectively. This performance notably surpasses several competitive methods, including ResNet101-RK [60] , RGTR&RPR [59], and SOLIDER [58]. The ResNet50 model also shown strong performance, achieving a mAP of 96.4%, Rank-1 accuracy of 99.1%, and Rank-5 accuracy of 99.5%.

Table 3. Results of models trained on UNI09

Models	mAP	Rank-1	Rank-5
OSNet_x1_0+TF (ours)	100	100	100
ResNet50+TF(ours)	99.9	100	100

In ***Table 3****,* it showcases the performance of our models when trained on the UNI09 dataset, a custom dataset designed for experimental analysis. The OSNet_x1_0 model achieved perfect scores, with a mAP, Rank-1, and Rank-5 accuracy of 100% across the board. The ResNet50 model also performed exceptionally well, with a mAP of 99.9% and perfect Rank-1 and Rank-5 accuracies of 100%.

5.1. Experiments

5.1.1. Model Training Configurations:

The OSNet_x1_0 model was configured using pre-trained weights from the ImageNet dataset, leveraging its generalization capabilities across diverse visual tasks. The Market1501 dataset comes pre-divided into training and testing sets. The OSNet_x1_0 model was trained on the training set, with image resizing to 256x128 pixels and various data augmentation methods, including random cropping, horizontal flipping, and normalization, were implemented. The training employed Adam optimizer and a learning rate set to 0.03, a train batch size of 300, and testing used a batch size of 80. The model was trained on a combined dataset of Market1501 and UNI09. The softmax loss function was used (7), and the training loop consisted of 50 epochs, focusing on the classifier layers for the first 10 epochs. The model training took 10 hours 57 minutes and 34 seconds. For the ResNet50 model, the training batch size was set to 400, with a testing batch size of 100. The training loop ran for 100 epochs in which 30 epochs was for classifier layer, with other parameters similar to those used for OSNet_x1_0. The model training for this took 29 hours 16 minutes and 28 seconds. A notable difference between the two models is their storage size. ResNet50, employing traditional 3x3 convolutional layers, has a significantly larger model size compared to OSNet_x1_0, which utilizes modern, lite 3x3 convolutional layers. Specifically, the OSNet_x1_0 model occupies 34.2 MB, whereas ResNet50 requires 305 MB. This difference results in a storage size gap of 270.8 MB, contributing to OSNet_x1_0's superior processing speed. The environment for model training and inference was set in PyTorch with CUDA 11.8 Installed in it.

5.1.2. Visual Ranking comparison:

The visual ranking comparison between the OSNet_x1_0 and ResNet50 models highlights the excellent output of OSNet_x1_0 in person ReID tasks. As shown in the visual ranking images, the OSNet_x1_0 model consistently ranks the correct matches with higher accuracy, evidenced by the higher number of green-bordered images, which indicate correct identifications.

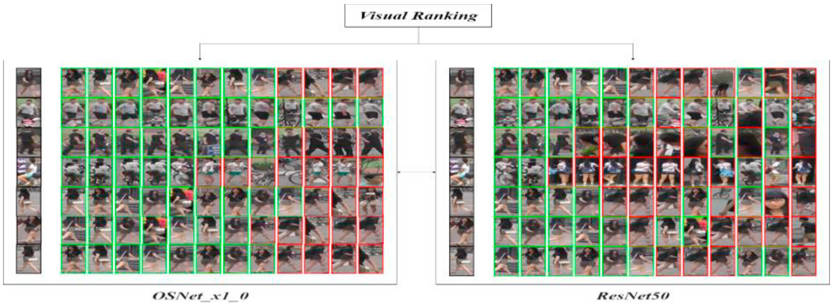

Fig 6. Visual ranking comparison between the models.

In contrast, the ResNet50 model displays a noticeable decrease in ranking precision, with a larger proportion of red-bordered images representing incorrect identifications as shown in ***Fig 6***. Overall, the visual ranking results clearly demonstrate that the OSNet_x1_0 model outperforms ResNet50 in this domain.

5.1.3. Activation Maps Generation:

Utilizing the technique described in [36], activation maps were generated to illustrate the areas of focus within the models during person ReID. It reveal that the OSNet_x1_0 model concentrates more precisely on key regions of the human body, which are crucial for distinguishing individuals.

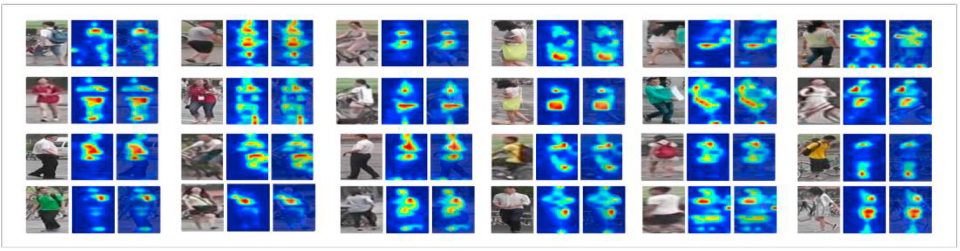

Fig 7. Activation and overlapped images generated from OSNet_x1_0 model.

These activation maps show that the model focuses on specific regions of the body, such as the torso, legs, and occasionally the head. The red and yellow areas indicate higher attention shown in ***Fig*** *7*, suggesting that the model is leveraging these regions to extract crucial features for distinguishing between individuals.

5.1.4. Actual Person ReID in Multi-Camera System:

The final output of the person ReID system demonstrates its ability to track and maintain consistent identities across multiple camera views, as shown in ***Fig 8***. Despite challenging conditions including changing lighting and angles, the system follows people well as they travel across many camera fields of view. Individual 1 (Green Bounding Box), This individual is consistently tracked across three different cameras. In Camera-1 (Back) and Camera-1 (Front), the person is identified as ID 3.

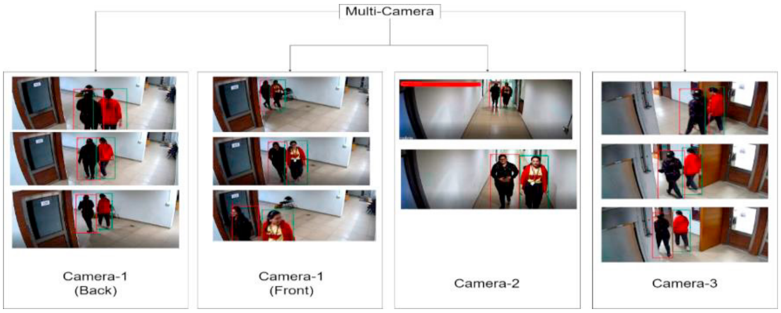

Fig 8. The Final output demonstration of Person ReID in Multi-Camera using OSNet_x1_0, the footage used here is taken from dataset UNI09.

The system maintains this ID when the individual appears in Camera-2 and Camera-3, despite changes in viewpoint and background. Individual **2** (Red Bounding Box), Similarly this individual also tracked through the same sequence and assigned the ID **5** and maintained throughout the fields. This ReID process is done using the state-of-the-art OSNet_x1_0 model. The system will firstly create a tracking combined video of multiple cameras. The tracking took around 394 seconds and the total frames of the all videos combined was 2756 frames. Each bounding box will be sent to transformer for resize and other preprocessing before it fed to ReID model.

Then the model will perform the compute distance as mentioned in ***Fig 5*** for ID assignment, after the process completes the final video will generate. Entire process took 492 seconds it may vary with different specifications, our computational specification were i7-12700, 32 DDR5 RAM and RTX 3060 12 GB (CUDA Enabled) for faster performance.

6. Conclusion

The study concludes with an overview of a multi-camera person ReID system that successfully tackles the challenges involved in tracking individuals from various camera angles. The proposed approach incorporates cutting-edge tracking and detection models, such as DeepSORT for MOT, a transformer model for bounding box analysis, and YOLOv8n for human detection. Furthermore, outstanding accuracy has been observed with the OSNet_x1_0 and ResNet50 models based on transfer learning, which have been trained on the Market1501 and custom UNI09 datasets simultaneously. The results, particularly the high mAP and Rank-1 accuracy attained on both datasets, show the efficiency of the suggested strategy in practical scenarios. Notably, the OSNet_x1_0 model was chosen due to its lightweight architecture as it only took 492 seconds to process 2765 frames, which provides high performance while being resource-efficient, making it an ideal choice for deployment in real-world multi-camera systems. The system's ability to maintain accurate identification across numerous camera views even under challenging conditions highlights its potential for wide applicability across various contexts.

7. References

[1] N. Narayan, N. Sankaran, D. Arpit, K. Dantu, S. Setlur and V. Govindaraju, "Person Re-identification for Improved Multi-person Multi-camera Tracking by Continuous Entity Association," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 566-572.

[2] Z. Zhun, L. Zheng, Z. Zheng, S. Li and Y. Yang, "Camera Style Adaptation for Person Re-identification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 5157-5166, 2018.

[3] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME--Journal of Basic Engineering,* vol. 82, no. Series-D, pp. 35-45, 1960.

[4] M. Bashar, S. Islam, K. K. Hussain, M. B. Hasan, A. B. M. A. Rahman and M. H. Kabir, "Multiple Object Tracking in Recent Times: A Literature Review," vol. abs/2209.04796, 2022.

[5] T. Diwan, G. Anirudh and J. V. Tembhurne, " Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications,* vol. 82, no. 6, pp. 9243-9275, 2023.

[6] A. Younis, L. Shixin, S. Jn and Z. Hai, "Real-Time Object Detection Using Pre-Trained Deep Learning Models MobileNet-SSD," in *ICCDE 2020: 2020 The 6th International Conference on Computing and Data Engineering*, 2020.

[7] C. K. Kumar and K. Rawal, "A Brief Study on Object Detection and Tracking," in *4th International Conference on Intelligent Circuits and Systems*, 2022.

[8] W. Rahmaniar and A. Hernawan, "Real-Time Human Detection Using Deep Learning on Embedded Platforms: A Review," *Journal of Robotics and Control (JRC),* vol. 2, no. 6, pp. 462-468, 2021.

[9] Srilaxmi, S. Kamath and V. Mayya, "Automated Human Detection in Images Using Deep Learning," *Tuijin Jishu/Journal of Propulsion Technology,* vol. 44, no. 5, pp. 1897-1902, 2023.

[10] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR,* vol. arXiv:1409.1556, 2015.

[11] {. Sivachandiran, K. Mohan and G. M. Nazer, "Deep Learning driven automated person detection and tracking model on surveillance videos," *Measurement: Sensors,* vol. 24, p. 100422, 2022.

[12] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *CoRR,* vol. abs/1804.02767, 2018.

[13] H. Hu, R. Hachiuma, h. Saito, Y. Takatsume and H. Kajita, "Multi-Camera Multi-Person Tracking and Re-Identification in an Operating Room," *Journal of Imaging,* vol. 8, no. 219, 2022.

[14] A. Specker and J. Beyerer, "ReidTrack: Reid-only Multi-target Multi-camera Tracking," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 5442-5452, 2023.

[15] K. Jungling, C. Bodensteiner and M. Arens, "Person re-identification in multi-camera networks".

[16] T. D'Orazio and C. Guaragnella, "A GRAPH-BASED SIGNATURE GENERATION FOR PEOPLE RE-IDENTIFICATION IN A MULTI-CAMERA SURVEILLANCE SYSTEM," *International conference on computer vision theory and applications,* 2012.

[17] R. Zhang, M. Li, X. Lv and L. Gao, *Journal of Physics: Conference Series,* vol. 2054, 2023.

[18] J. K. KANG, T. M. HOANG and K. R. PARK, "Person Re-Identification Between Visible and Thermal Camera Images Based on Deep Residual CNN Using Single Input," *IEEE Access,* vol. 7, pp. 57972-57984.

[19] T. Zhang, L. Xie, L. Wei, Y. Zhang, B. Li and Q. Tian, "Single Camera Training for Person Re-identification," *CoRR,* vol. arXiv:1909.10848v1 [cs.CV], 2019.

[20] Y. Du, C. Lei, Z. Zhao, Y. Dong and F. Su, "Video-based Visible-Infrared Person Re-Identification with Auxiliary Samples," *IEEE Transactions on Information Forensics and Security,* vol. 19, pp. 1313-1325, 2024.

[21] J. Shi, X. Yin, Y. Chen, Y. Zhang, Z. Zhang, Y. Xie and Y. Qu, "Multi-Memory Matching for Unsupervised Visible-Infrared Person Re-Identification," in *arXiv*, 2024.

[22] S. Saber, S. Meshoul, K. Amin, P. Pławiak and M. Hammad, "A Multi-Attention Approach for Person Re-Identification Using Deep Learning," *Sensors,* vol. 23, no. 7, p. 3678, 2023.

[23] M. {Tahir and S. Anwar, "Transformers in Pedestrian Image Retrieval and Person Re-Identification in a Multi-Camera Surveillance System," *Applied Sciences,* vol. 11, no. 19, p. 9197, 2021.

[24] F. Cunico and M. Cristani, "Multi-Camera Industrial Open-Set Person Re-Identification and Tracking," *arXiv,* vol. 2409.03879, 2024.

[25] L. iang, M. Zhang, Z. Gao, Y. Li and L. Chai, "A multi-camera person tracking and re-identification system based on edge computing," in *Second International Conference on Electronic Information Engineering, Big Data, and Computer Technology EIBDCT 2023*, 2023, p. 1264230.

[26] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu and X. Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," *CoRR,* vol. abs/2110.06864, 2021.

[27] J. François, S. Wang, R. State and T. Engel, "BotTrack: Tracking Botnets Using NetFlow and PageRank," in *NETWORKING 2011*, 2011.

[28] Y. Zhang, C. Wang, X. Wang, W. Zeng and W. Liu, "A Simple Baseline for Multi-Object Tracking," *CoRR,* vol. abs/2004.01888, 2020.

[29] N. Wojke, A. Bewley and D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," *CoRR,* vol. abs/1703.07402, 2017.

[30] A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft, "Simple Online and Realtime Tracking," *CoRR,* vol. abs/1602.00763, 2016.

[31] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang and Q. Tian, "MARS: A Video Benchmark for Large-Scale Person Re-Identification," in *Computer Vision – ECCV 2016*, 2016.

[32] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable Person Re-Identification: A Benchmark," in *ICCV 2015*, 2015.

[33] K. Zhou and T. Xiang, "Torchreid: {A} Library for Deep Learning Person Re-Identification in Pytorch," *CoRR,* vol. abs/1910.10093, 2019.

[34] M.-Y. Huh, P. Agrawal and A. A. Efros, "What makes ImageNet good for transfer learning?," *CoRR,* vol. abs/1608.08614, 2016.

[35] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung and M. A. Azim, "Transfer learning: a friendly introduction," *Journal of Big Data ,* vol. 9, no. 1, 2022.

[36] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Omni-Scale Feature Learning for Person Re-Identification," *CoRR,* vol. arXiv:1905.00953v6 [cs.CV], 2019.

[37] K. Zhou, Y. Yang, A. Cavallaro and T. Xiang, "Learning Generalisable Omni-Scale Representations for Person Re-Identification," *TPAMI,* 2021.

[38] M. Safaldin, N. Zaghden and M. Mejdoub, "An Improved YOLOv8 to Detect Moving Objects," *IEEE Access,* vol. 12, pp. 59782-59806, 2024.

[39] A. R. o. Y. a. I. Advancements, "Sohan, Mupparaju; Sai Ram, Thotakura;Rami Reddy, Ch. Venkata;," in *Data Intelligence and Cognitive Informatics*, Singapore, Springer Nature Singapore, 2024, pp. 529-545.

[40] J. R. Terven and . M. Cordova-Esparza, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction,* vol. 5, no. 4, p. 1680–1716, 2023.

[41] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen and J.-W. Hsieh, "CSPNet: A New Backbone that can Enhance Learning Capability of CNN," *CoRR,* vol. abs/1911.11929, 2019.

[42] D. Li, X. Chen, Z. Zhang and K. Huang, "Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification," *CoRR,* vol. abs/1710.06555, 2017.

[43] A. Gupta, P. Pawade and R. balakrishnan, "Deep Residual Network and Transfer Learning-based Person Re-identification," *Intelligent Systems with Applications,* 2022.

[44] L. Wei, S. Zhang, H. Yao, W. Gao and Q. Tian, "GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval," *CoRR,* vol. abs/1709.04329, 2017.

[45] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang, "Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline)," *CoRR,* vol. abs/1711.09349, 2017.

[46] N. Martinel, G. L. Foresti and C. Micheloni, "Aggregating Deep Pyramidal Representations for Person Re-Idenfitication," in *International Conference on Computer Vision and Pattern Recognition Workshops (CVPR) 2019*.

[47] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang and Y. Wei, "Circle Loss: A Unified Perspective of Pair Similarity Optimization," *CoRR,* vol. abs/2002.10857, 2020.

[48] J. Li and X. Gong, "Prototypical Contrastive Learning-based CLIP Fine-tuning for Object Re-identification," no. arXiv, 2023.

[49] D. Fu, D. Chen, H. Yang, J. Bao, L. Yuan, L. Zhang, H. Li, F. Wen and D. Chen, "Large-Scale Pre-training for Person Re-identification with Noisy Labels," no. arXiv, 2022.

[50] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang and J. Kautz, "Joint Discriminative and Generative Learning for Person Re-identification," *CoRR,* vol. abs/1904.07223, 2019.

[51] X. Zang, G. Li, W. Gao and X. Shu, "Learning to Disentangle Scenes for Person Re-identification," *CoRR,* vol. abs/2111.05476, 2021.

[52] Y. Chen, H. Wang, X. Sun, B. Fan and C. Tang, "Deep Attention Aware Feature Learning for Person Re-Identification," *Pattern Recognition,* vol. 126, p. 108567, 2022.

[53] Q. Wang, X. Qian, B. Li, Y. Fu and x. xue, "Rethinking Person Re-identification from a Projection-on-Prototypes Perspective".

[54] V. Somers, C. D. Vleeschouwer and A. Alahi, "Body Part-Based Representation Learning for Occluded Person Re-Identification," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV*, IEEE, 2023.

[55] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg and G. Rigoll, "Lightweight Multi-Branch Network for Person Re-Identification," *CoRR,* vol. abs/2101.10774, 2021.

[56] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, W.-S. Zheng and X. Sun, "Viewpoint-Aware Loss with Angular Regularization for Person Re-Identification," *CoRR,* vol. abs/1912.01300, 2019.

[57] G. Wang, J.-H. Lai, P. Huang and X. Xie, "Spatial-Temporal Person Re-identification," *CoRR,* vol. abs/1812.03282, 2018.

[58] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin and X. Sun, "Beyond Appearance: a Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks," 2023.

[59] Y. Gong, L. Huang and L. Chen, "Eliminate Deviation with Deviation for Data Augmentation and a General Multi-modal Data Learning Method," 2022.

[60] D. Fu, D. Chen, J. Bao, H. Yang, L. Yuan, L. Zhang, H. Li and D. Chen, "Unsupervised Pre-training for Person Re-identification," *CoRR,* vol. abs/2012.03753, 2020.