# Multiview Pedestrian Detection with Multi Pedestrian Consistency Loss

Myeongjun Kim[1,2][0000−0001−9500−8491]

[1] DeepingSource, Republic of Korea
[2] POSTECH, Pohang, Republic of Korea
myeongjun@postech.ac.kr

**Abstract.** Recent advances in deep learning have led to significant breakthroughs across various fields, including both image and video analysis. However, single-image and single-video approaches have inherent limitations in their viewpoints. In particular, surveillance camera analysis often relies on individual cameras, even when multiple cameras cover the same location, limiting the scope of information available from a single viewpoint. To address these limitations, we focus on pedestrian detection and tracking in multi-camera settings, which can help mitigate issues of person occlusion and missed detections that are common in single-view detection and tracking tasks. In tasks involving multiview inputs, inter-view correlations offer valuable contextual information. To utilize this advantage, we propose a novel approach comprising a consistency loss and a feature aggregation module designed to enhance multiview correlations. First, we introduce a Multi-Pedestrian Consistency Loss to maintain feature coherence for individuals appearing across different views. Second, we propose the Multi-Camera Feature Aggregation module, which captures broader contextual information from surrounding areas using a large receptive field. Our proposed methods are evaluated on the WildTrack and MultiViewX datasets, where it achieves state-of-the-art performance across key metrics in both detection and tracking tasks.

**Keywords:** Multi-View Pedestrian Detection & Tracking, Cross-View Consistency, Feature Aggregation.

## 1 Introduction

Deep learning has profoundly transformed the fields of image and video analysis, achieving remarkable success in tasks such as object detection [3, 25] and tracking [1, 22]. However, methods relying on single images or single video streams are inherently constrained by their viewpoints, leading to challenges like occlusions and missed detections, especially in crowded environments. This limitation is particularly evident in surveillance systems, where multiple surveillance cameras may cover the same area, yet analysis is often conducted using data from a single camera. The reliance on a single viewpoint not only restricts the richness of the available information but also compromises the robustness of detection

and tracking algorithms. To address these challenges, multi-camera multi-view approaches have gained attention, leveraging information from multiple perspectives to mitigate occlusions and improve detection accuracy. Multi-Pedestrian detection and tracking in such settings enhance scene understanding, which is crucial for applications like surveillance, crowd monitoring, and autonomous systems. Integrating data from multiple cameras makes it possible to construct a more comprehensive representation of the environment, leading to better performance in detection and tracking tasks.

Fig. 1: Illustration of the Multi-Pedestrian Consistency Loss (MPCL). The four provided images (top-left, top-right, bottom-left, bottom-right) represent camera views from different angles at the same time $t$. Each colored box (Red, Light Blue, Green, Dark Blue, Purple) denotes the bounding box of a person in each view. By applying consistency loss, our model is trained to maintain feature consistency across multiple camera views, ensuring that the same person's features remain coherent despite variations in perspective.

Existing multiview pedestrian detection and tracking approaches [30, 16, 15] typically extract features using backbone networks [14, 10, 28], project these features through lifting methods [16], and ultimately generate an occupancy map. However, previous methods face two key limitations. First, while multiview inputs are utilized, inter-view correlations are not exploited. In human perception, constructing an occupancy map inherently involves identifying the locations of the same person across multiple views and leveraging surrounding context to establish correlations between views. Additionally, the model that effectively captures the spatial correlations in a multiview setting can also demonstrate strong performance in downstream tasks, such as person attribute analysis [20] (e.g., age, gender, etc.) or person Re-identification [31]. Second, current methods difficult to incorporate surrounding context when aggregating projected features after

lifting, which limits the overall accuracy and contextual consistency of the aggregation. To address the first issue, we propose an MPCL: Multi-Pedestrian Consistency Loss that promotes feature consistency for individuals across views, guiding the model to learn meaningful inter-view correlations. For the second issue, we introduce an MCFA: Multi-Camera Feature Aggregation module. While prior methods rely on simple convolutional blocks for feature aggregation after lifting, our MCFA module enables aggregation that accounts for global context, thus enhancing feature integration. By employing the proposed MPCL and MCFA module, our network achieves state-of-the-art (SOTA) performance on the WildTrack and MultiViewX datasets, demonstrating its effectiveness in addressing the limitations of prior approaches in multiview pedestrian detection and tracking.

Our contributions are summarized as follows:

- We introduce an MPCL that promotes feature consistency across camera views, improving the association of individuals in multi-camera multiview settings.
- We propose an MCFA module that leverages a large receptive field during the multiview feature aggregation stage, enabling the model to consider the global context.
- We evaluated the MultiViewX [16] and WildTrack [4] datasets, achieving state-of-the-art (SOTA) performance on various metrics (detection and tracking).

## 2 Related Works

**Multiview Pedestrian Detection** Object detection methods struggle with occlusion in crowded scenes due to reliance on a single viewpoint. Multiview pedestrian detection addresses this limitation by leveraging multiple camera views and projecting them onto a common ground plane, allowing occluded pedestrians in one view to be visible in another. Various projection methods have been explored, including perspective projection [16], deformable attention [19], and depth-based approaches. Among these, perspective projection is widely used due to its efficiency and effectiveness.

**Multiview tracking** aims to maintain consistent identities of pedestrians across multiple camera views. Traditional approaches rely on appearance features and geometric constraints, while recent methods employ deep learning for cross-view identity matching. However, significant viewpoint variations and occlusions remain challenging.

**Similarity learning** aims to learn representations that are invariant to transformations. Self-supervised methods such as SimCLR [6] and MoCo [12] have demonstrated strong performance by leveraging augmented views of the same image. However, these approaches often operate at the image level, limiting their ability to distinguish between individual instances. Instance-level methods address this by focusing on object-specific representations, but handling viewpoint variations remains challenging.

**Multi-Camera Feature Aggregation** Feature aggregation is crucial in multi-camera pedestrian detection for combining information across views. Traditional methods rely on simple fusion strategies such as max-pooling or averaging [4], which often fail to capture global context. Recent approaches leverage attention mechanisms and transformer-based architectures [19] to improve multi-view feature fusion. However, these methods can introduce unnecessary global interactions and increase computational complexity.

## 3 Methods

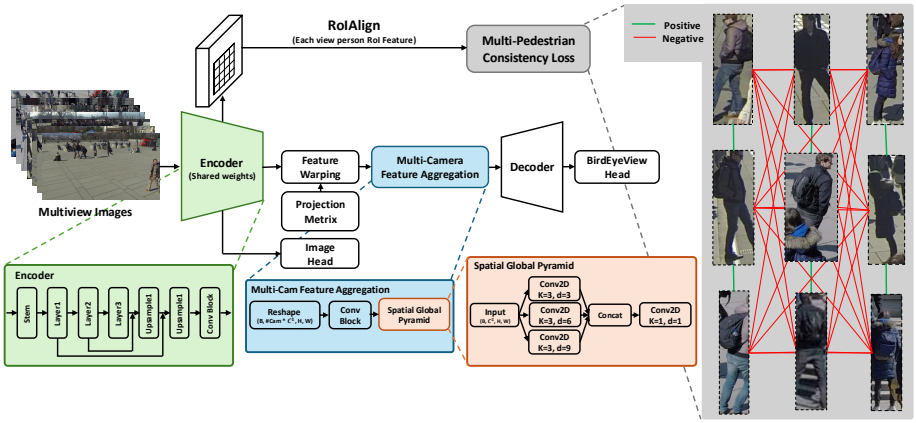


Fig. 2: Overall architecture of the proposed network. The network takes multiview images as input and extracts features through the Encoder (backbone, shown in Green). Person patch features are then obtained using RoIAlign and are used in the MPCL (Gray) training. Additionally, the encoder output features are fed into the MCFA module (Light Blue) to obtain features with global context awareness. Finally, the Decoder network generates the occupancy map. In subfigure (B), positive and negative features are distinguished, where features with the same person-ID across multiple cameras are considered positive (Green line), and all others are treated as negative (Red line) for training.

In this section, we will deep dive into the proposed loss and trainable projection method. Before introducing the module, let's talk about multiview pedestrian detection. The multiview pedestrian detection task receives multiple images as input, extracts features through the backbone network, and projects the extracted features onto a single ground plane using projection matrices to extract projected features. The goal is to then draw an occupancy map. We propose a consistency loss that can help with detection and tracking, and enhance multi-camera view aggregation by adding a trainable Multi-Camera Feature Aggregation module (MCFA) to the traditional projection method (perspective transformation [16]) that has been widely used in previous networks. [16, 30]

The overall network works as follows. Given $N$ camera views, each view captures an image of size $H \times W \times C$. These $N$ images are stacked to form an input $x$ ($x$ has size $(N \times C \times H \times W)$[3]). The input multi-view images are first processed by a 2D image backbone network, which extracts image features for each view. These extracted image features serve as inputs to two distinct modules:

1. Multi-Pedestrian Consistency Loss (MPCL): To enforce consistency across views, person patch features are extracted from the extracted view features using RoIAlign [13]. These features are then used to compute the person consistency loss, ensuring coherence in person appearance across multiple views.
2. Multi-Camera Feature Aggregation (MCFA): The extracted features from different views are projected onto a common space using pre-defined projection matrices. The projected features are then aggregated and processed through the MCFA module to generate a single, unified occupancy map.

This approach enables effective fusion of multi-view information, leveraging both view-specific consistency and spatial aggregation to enhance scene understanding.

### 3.1 Multi-Pedestrian Consistency Loss

For multiview detection and tracking, it is important to consistently recognize the same person in the spatial and temporal axes. When approaching the tracking problem from the perspective of human perception, even if the images are taken from different views, the tracking task is performed by finding the features of the same person in each view and consistently considering them as one. In a similar way, we proposed a Person Consistency Loss that can maintain consistency between person patch features in the spatial axes of different views in order to train similarly to human perception.
The instance consistency loss was proposed in Matching Anything by Segmenting Anything (MASA) [18]. However, the loss provided in MASA has a difficulty in that it is trained as negative when the same person exists in the batch. Therefore, in our setting, as shown in the figure above, multiple person patch features can be generated for each view, and if they have the same person-id, they are trained as positive, and the rest are learned as negative. Person patch features for person consistency loss are generated by the RoIAlign on the output features of the backbone network. (Each feature has dimensions of $C^1 \times H \times W$, where $C^1$ is the $out\_channel$ of the backbone network.)

$$L_C = -\sum_{q \in Q} log \frac{\sum_{q^+ \in Q^+} e^{\frac{sim(q,q^+)}{\tau}}}{\sum_{q^+ \in Q^+} e^{\frac{sim(q,q^+)}{\tau}} + \sum_{q^- \in Q^-} e^{\frac{sim(q,q^-)}{\tau}}} \tag{1}$$

Here, $q^+$ and $q^-$ denote the positive and negative samples to $q$, respectively. Unlike the original consistency loss, in our task, we can obtain person patch features

[3] Since our method utilizes a 2D backbone (ResNet-18), the final input to the backbone network is reshaped to $((B * N) \times C \times H \times W)$.

for each view, and thus there are as many patch features with the same person-id as the number of cameras. Therefore, we repurpose the original consistency loss to train to attract positive person patches and repel negative person patches. MPCL is multiplied by $\lambda_{MPCL}$ and added to the original loss function. For the WildTrack dataset, $\lambda_{MPCL}$ is set to 0.7, while for the MultiViewX dataset, it is set to 0.5.

### 3.2 Multi-Camera Aggregation Module

In previous works [16, 15, 30], occupancy maps are generated by first extracting image features for each view, with feature dimensions of $(B * N) \times C \times H \times W$. These features are projected onto a common space using projection matrices, then concatenated across views, resulting in a tensor of shape $B \times (C * N) \times H \times W$. While effective, simple concatenation limits the global receptive field, as 3x3 convolutions applied to the concatenated features only capture a small receptive field, potentially missing broader spatial contexts. To address this limitation, we propose a Multi-Camera Feature Aggregation Module (MCFA) that incorporates convolutions with varying dilation sizes, thereby enhancing the global receptive field and effectively aggregating information from multiple views. Our method restructures the concatenated projected features to better capture multi-view context while maintaining spatial alignment across views. In our approach, we reshape the feature tensor along the camera axis to form $x \in \mathbb{R}^{B \times (C*N) \times H \times W}$, and then reduce this feature dimensionality using a simple convolutional block, yielding $x' \in \mathbb{R}^{B \times C_{proj} \times H \times W}$. This transformed feature, $x'$, is then fed into our Spatial Global Pyramid Module, which consists of convolutional blocks with various dilation sizes. By employing multiple dilation rates, we capture a range of receptive fields, preserving both local and global contextual information.

In our approach, features extracted from each view are reshaped to a tensor of size $B \times (N \cdot C^1) \times H \times W$, followed by aggregation using the Spatial Global Pyramid. This mechanism effectively captures the global context. However, the high computational cost poses a significant challenge. To address this, we apply a convolutional block for channel reduction before performing aggregation. Comparing the computational complexity before and after channel reduction, the complexity without reduction is given by:

$$O((C^1 \cdot N) \cdot C^2 \cdot H \cdot W)$$

whereas the complexity with reduction is:

$$O((C^1 \cdot N) \cdot C^2 \cdot H \cdot W + C^2 \cdot C^2 \cdot H \cdot W)$$

which results in an approximate 89% reduction in computational cost. This demonstrates the efficiency of the proposed approach in reducing computational overhead while preserving global contextual information.
Additionally, through experimental analysis, we observed that using excessively large dilation sizes introduced noise, particularly from distant background regions or irrelevant human features, leading to a performance decline. This indicates that for accurate occupancy mapping, overly distant context can act as

noise, thus an optimal selection of dilation sizes is essential for balancing local and global feature representation in MCFA.

## 4 Experiments

### 4.1 Datasets

**WildTrack** dataset [4] is a real-world dataset consisting of 400 synchronized frames. Each camera operates at a resolution of 1080 × 1920 pixels and captures frames at 2 frames per second (FPS). A total of 7 cameras cover a single public square, capturing overlapping views of a 12 × 36 m area, consistent with previous works. Ground-plane annotations are provided on a 480 × 1440 grid, corresponding to 2.5 cm grid cells. On average, each frame contains 20 pedestrians, observed by 3.74 cameras. Due to the high pedestrian density in a small area, the dataset exhibits significant occlusion. Additionally, since all cameras focus on the same space, they provide overlapping views, making the dataset somewhat limited as a real-world setting.

**MultiViewX** dataset [16] is a synthetic dataset specifically designed for multiview pedestrian detection tasks, covering an area of 16 meters by 25 meters, slightly smaller than the WildTrack dataset. Similar to WildTrack, the ground plane is divided into a grid with dimensions of 640 × 1000 cells, each cell representing a 2.5cm square to enable fine-grained annotation. MultiViewX includes 6 cameras with overlapping fields of view, each capturing images at a resolution of 1080 × 1920 pixels. Annotations are provided for 400 frames recorded at 2 fps, matching the temporal resolution of WildTrack, and each scene location is, on average, covered by approximately 4.41 cameras, providing high visibility across different views. As a synthetic dataset, MultiViewX allows for flexible scenario configurations with freely available annotations. In its default configuration, the dataset includes an average of 40 pedestrians per frame, effectively doubling the crowd density of WildTrack and offering a more challenging setting for evaluating detection and tracking performance in high-density pedestrian environments.

### 4.2 Evaluation Metrics

Unlike conventional monocular-view detection, multiview detection generates a ground plane occupancy map, which is evaluated based on the distance to the ground-truth occupancy. Our evaluation follows the methodology outlined in previous works [16]. For Multi-Target Multi-Camera tracking, we adopt the 2D Bird's-eye view center points protocol, as used in prior research [30], to ensure consistency in tracking evaluation across different perspectives.

**Detection Metrics** A pedestrian detection is considered a true positive if it is within a radius of $r = 0.5$ meters from the ground-truth position, roughly corresponding to the average radius of a human body. In the detection task, accurate object detection is crucial. Therefore, we consider Multiple Object Detection Accuracy (MODA) as the most important performance metric, as it evaluates

Table 1: Pedestrian detection results on the WildTrack and MultiViewX datasets. **Bold** values indicate the best overall results.

| Model | Lifting Method | WildTrack | | | | MultiViewX | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MODA | MODP | Precision | Recall | MODA | MODP | Precision | Recall |
| DeepMCD [5] | Learned | 67.8 | 64.2 | 85 | 82 | 70.0 | 73.0 | 85.7 | 83.3 |
| Deep-Occlusion [2] | Learned | 74.1 | 53.8 | 95 | 80 | 75.2 | 54.7 | 97.8 | 80.2 |
| MVDet [16] | Persp. Proj. | 88.2 | 75.7 | 94.7 | 93.6 | 83.9 | 79.6 | 96.8 | 86.7 |
| SHOT [27] | Persp. Proj. | 90.2 | 76.5 | 96.1 | 94.0 | 88.3 | 82.0 | 96.6 | 91.5 |
| 3DROM [24] | Persp. Proj. | 91.2 | 76.9 | 95.9 | 95.3 | 90.0 | 83.7 | 97.5 | 92.4 |
| MVDeTr [15] | Persp. Proj. | 91.5 | **82.1** | 97.4 | 94.0 | 93.7 | 91.3 | 99.5 | 94.2 |
| EarlyBird [29] | Persp. Proj. | 91.2 | 81.8 | 94.9 | 96.3 | 94.2 | 90.1 | 98.6 | 95.7 |
| MVTT [17] | Persp. + RoI | 94.1 | 81.3 | **97.6** | 96.5 | 95.0 | **92.8** | 99.4 | 95.6 |
| TrackTacular [30] | Persp. Proj. | 91.8 | 79.8 | 96.2 | 95.6 | 95.9 | 89.2 | 99.5 | 96.4 |
| **Ours** | Persp. Proj. | **94.43** | 81.01 | 97 | **97.48** | **96.52** | 87.65 | **99.5** | **97.1** |

overall performance by accounting for both false positives (FP) and false negatives (FN). Next, we prioritize Precision and Recall as the second most important metrics. Finally, we consider Multiple Object Detection Precision (MODP) as the third key metric. While localization accuracy is important, the primary focus in detection tasks is ensuring that objects are correctly detected.

**Tracking Metrics** We report standard Multiple Object Tracking (MOT) metrics alongside identity-aware metrics. A predicted track is considered a positive match with a ground-truth track if it falls within a radius of $r = 1$ meter. In the tracking task, the ability to maintain object identities under occlusion and overall tracking accuracy are critical. Therefore, we consider IDF1 and Multiple Object Tracking Accuracy (MOTA) as the primary performance metrics. Additionally, we utilize MOTP, MT, and ML as supplementary metrics to assess localization accuracy and overall tracking performance, enhancing the reliability of the results.

### 4.3 Implementation Details

Following the methodologies proposed in [30], we apply data augmentation to the RGB input images by performing random resizing and cropping within a scaling factor range of [0.8, 1.2]. Applying random cropping may result in person patches being cut off, making them unrecognizable as human figures. Therefore, unusable patches were removed, ensuring that only meaningful person patches were used for MPCL training. For MPCL, we use $512 \times 14 \times 14$ person patches with RoIAlign [13]. For training, we utilized a one-cycle learning rate scheduler across both datasets, with a batch size of 1. The Adam optimizer was employed on the WildTrack and MultiViewX datasets with a learning rate of 0.0015. To ensure stable training and effective utilization of batch statistics, we accumulate gradients over multiple iterations before updating the network weights, effectively achieving an equivalent batch size of 8. Both the encoder and decoder networks are initialized with weights pre-trained on the ImageNet-1K dataset [9].

Table 2: Pedestrian tracking results on the WildTrack and MultiViewX datasets. **Bold**: Values indicate the best overall results. (↑): Higher values indicate better performance. (↓): Lower values indicate better performance.

| Method | WildTrack | | | | |
|---|---|---|---|---|---|
| | **IDF1↑** | **MOTA↑** | **MOTP↑** | **MT↑** | **ML↓** |
| KSP-DO [4] | 73.2 | 69.6 | 61.5 | 28.7 | 25.1 |
| KSP-DO-ptrack [4] | 78.4 | 72.2 | 60.3 | 42.1 | 14.6 |
| GLMB-YOLOv3 [23] | 74.3 | 69.7 | 73.2 | 79.5 | 21.6 |
| GLMB-DO [23] | 72.5 | 70.1 | 63.1 | **93.6** | 22.8 |
| DMCT [32] | 77.8 | 72.8 | 79.1 | 61.0 | 4.9 |
| DMCT Stack [32] | 81.9 | 74.6 | 78.9 | 65.9 | 4.9 |
| ResT [7] | 86.7 | 84.9 | 84.1 | 87.8 | 4.9 |
| EarlyBird [29] | 92.3 | 89.5 | 86.6 | 78.0 | 4.9 |
| MVFlow [11] | 93.5 | 91.3 | - | - | - |
| TrackTacular [30] | 94.2 | 89.6 | 81.7 | 87.8 | 4.9 |
| **Ours** | **96.01** | **92.75** | **88.36** | 87.8 | **4.9** |
| | **MultiViewX** | | | | |
| | **IDF1↑** | **MOTA↑** | **MOTP↑** | **MT↑** | **ML↓** |
| EarlyBird [29] | 82.4 | 88.4 | 86.2 | 82.9 | 1.3 |
| TrackTacular [30] | 84.2 | 91.4 | **86.7** | 85.5 | 2.6 |
| **Ours** | **86.61** | **92.57** | 84.96 | **90.79** | **1.3** |

### 4.4 Comparison to Previous Works

Quantitative results were obtained on the WildTrack and MultiViewX datasets. Detection metrics are reported in Table 1 while tracking metrics are presented in Table 2.

*Detection Results.* For the WildTrack dataset, our approach achieves state-of-the-art (SOTA) performance on two out of four metrics. Notably, our model improves MODA by +**0.33** compared to the MVTT [17] and by +**2.63** over the TrackTacular [30]. On the MultiViewX dataset, our model attains SOTA performance on three out of four metrics, with a +**0.62** improvement over TrackTacular. Overall, MODA, which measures the accuracy of detected objects by accounting for false positives and false negatives, and Recall, which represents the proportion of detected objects among existing ones, are essential for pedestrian detection. Our model demonstrates SOTA performance on both metrics.

*Tracking Results.* In terms of tracking, our model achieves SOTA performance on four out of five metrics on the WildTrack dataset, with an improvement of +**1.81** in IDF1 over the previous SOTA, TrackTacular, and +**1.45** in MOTA

Table 3: Effectiveness of the Multi-Pedestrian Consistency Loss (MPCL) and Multi-Camera Feature Aggregation (MCFA) module.

| Method | WildTrack | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **MODA** | **MODP** | **Precision** | **Recall** | **IDF1** | **MOTA** | **MOTP** | **MT** | **ML** |
| Baseline (TrackTacular) | 91.8 | 79.8 | 96.2 | 95.6 | 94.2 | 89.6 | 81.7 | 87.8 | 4.9 |
| + MPCL | 94.01 | 79.05 | **97.15** | 96.84 | 95.2 | 91.7 | 86.5 | 85.36 | 4.9 |
| + MCFA | **94.43** | **81.01** | 97 | **97.48** | **96.01** | **92.75** | **88.36** | **87.8** | **4.9** |
| | **MultiViewX** | | | | | | | | |
| | **MODA** | **MODP** | **Precision** | **Recall** | **IDF1** | **MOTA** | **MOTP** | **MT** | **ML** |
| Baseline (TrackTacular) | 95.9 | 89.2 | 99.5 | 96.4 | 84.2 | 91.4 | **86.7** | 85.5 | 2.6 |
| + MPCL | 96.05 | **89.3** | 99.25 | 96.8 | 85.7 | **92.3** | 85 | **89.4** | 1.3 |
| + MCFA | **96.52** | 88.5 | **99.5** | **97.1** | **86.2** | 91.23 | 85.3 | 86.8 | **1.3** |

over MVFlow[11]. Similarly, on the MultiViewX dataset, our approach achieves SOTA on four out of five metrics, with IDF1 showing a $+\mathbf{2.41}$ improvement over TrackTacular and MOTA showing a $+\mathbf{1.17}$ improvement. In pedestrian tracking, IDF1, which assesses the consistent ID matching between tracked objects and real objects, and MOTA, which evaluates the overall tracking errors, are critical metrics. Our model achieves SOTA performance on these metrics across both datasets.

### 4.5 t-SNE Visualization & Quantitative Results

The t-SNE [21] plot below visualizes features obtained by applying RoIAlign to the person patches of each view, using t-SNE for dimensionality reduction. First, when comparing the WildTrack dataset with and without the Multi-pedestrian Consistency loss, we observe that in the Multi-pedestrian Consistency loss setting, clusters are better preserved for each view. Examining the image closely, each color represents a different person ID, but some overlap in colors is present. Notably, clusters of the same color do not merge into a single cluster; instead, each view's cluster remains distinct, with some separation between them. This occurs because the features observed from different views vary significantly. For example, the front and back views of a person exhibit entirely different features. Consequently, these two views cannot form a single cluster. While clusters from different views are slightly separated, they still appear to be part of a larger group when considered as a whole. Second, in the MultiViewX dataset, we observe that clusters are formed as intended when using Multi-pedestrian Consistency loss. Additionally, despite the number of person IDs being higher than in the WildTrack dataset, the clusters are appropriately structured, demonstrating the effectiveness of the loss function in maintaining consistency across views. Visualization ambiguity arises when multiple feature points share the same color due to the limited number of available colors relative to the number of person IDs. As a result, multiple clusters may be represented by the same color, leading to potential misinterpretation. Finally, we bring in numerical metrics, the Silhouette Coefficient (SC) [26] and the Davies-Bouldin Index (DB Index) [8],

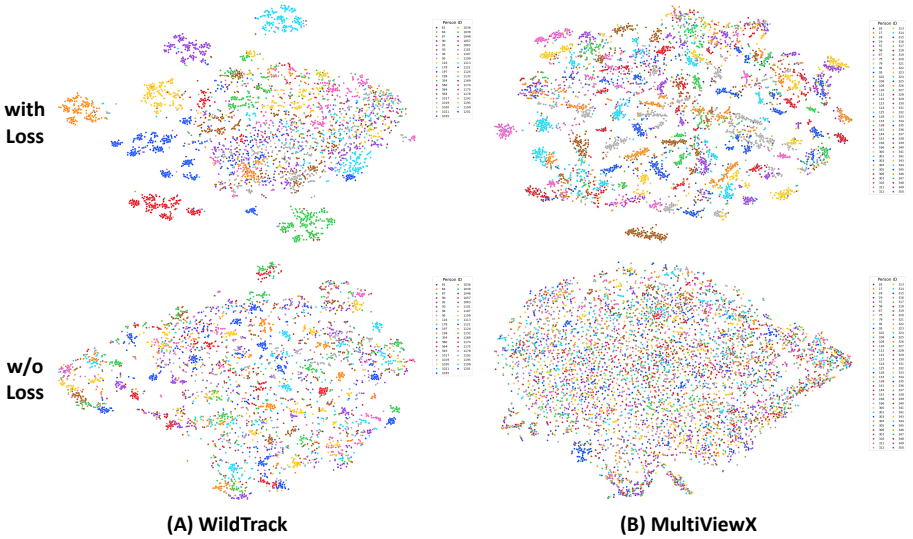


Fig. 3: t-SNE plots of person patch features on the WildTrack and MultiViewX datasets, illustrating the effect of consistency loss. The plots compare the distribution of person patch features with and without the application of consistency loss. By employing consistency loss, the features for each person become more coherent, resulting in improved feature clustering across multiple views.

which measure clustering quality for each dataset. The SC evaluates clustering quality by comparing an object's similarity to its own cluster (cohesion) versus other clusters (separation). It ranges from -1 to 1. The DB Index is a metric used to evaluate the quality of clustering in unsupervised learning. It measures the average similarity ratio of each cluster with its most similar cluster, with lower values indicating better clustering. The index considers both intra-cluster dispersion (how tightly clustered the points are within a cluster) and inter-cluster separation (how distinct or far apart the clusters are from each other). In table 4, we can see that with Multi-pedestrian Consistency loss, clusters are better formed compared to the without loss setting.

Table 4: Comparison of our Model's t-SNE quantitative result performance on the WildTrack and MultiViewX datasets. Abbreviations are the following. SC: Silhouette Coefficient [26], DB Index: Davies-Bouldin Index [8]

| Model | Wildtrack | | MultiviewX | |
|---|---|---|---|---|
| | SC ↑ | DB Index ↓ | SC ↑ | DB Index ↓ |
| Ours (**w loss**) | **0.051** | **10.48** | **-0.286** | **5.92** |
| Ours (**w/o loss**) | -0.278 | 72.472 | -0.341 | 60.58 |

Table 5: Comparison of our Model's performance against previous lifting methods on the WildTrack and MultiViewX datasets.

| Model | Lifting Method | WildTrack | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **MODA** | **MODP** | **Precision** | **Recall** | **IDF1** | **MOTA** | **MOTP** | **MT** | **ML** |
| TrackTacular [30] | Persp. Proj. | 91.8 | 79.8 | 96.2 | 95.6 | 94.2 | 89.6 | 81.7 | 87.8 | 4.9 |
| **Ours** | Persp. Proj. | **94.64** | **81.5** | **97.9** | **96.7** | **96.2** | **93.3** | **87.1** | **87.8** | **4.9** |
| TrackTacular [30] | Bilin. Sampl. | 92.1 | 76.2 | 97 | 95.1 | 95.3 | 91.8 | 85.4 | 87.8 | 4.9 |
| **Ours** | Bilin. Sampl. | **93.4** | **81.1** | **97.13** | **96.2** | **95.78** | **92.54** | **86.4** | **87.8** | **4.9** |
| TrackTacular [30] | Depth Splat. | 93.2 | 77.5 | **97.3** | 95.8 | 93.6 | 90.2 | 85.4 | **87.8** | 4.9 |
| **Ours** | Depth Splat. | **93.3** | **83** | 97.13 | **96.11** | **96.1** | **92.8** | **87.84** | 85.4 | **4.9** |
| TrackTacular [30] | Deform. Attn. | 78.4 | 73.1 | 93.8 | 84 | 88 | 82.2 | 78.9 | **75.6** | 4.9 |
| **Ours** | Deform. Attn. | **82.03** | **78.9** | **96.43** | **85.2** | **88.7** | **83.4** | **83.4** | 73.2 | **4.9** |

### 4.6 Module Contributions

The contributions of the two proposed modules are presented in Table 3. On the WildTrack dataset, the Multi-Pedestrian Consistency Loss demonstrates a dominant improvement in performance. Additionally, the Multi-Camera Feature Aggregation (MCFA) module yields significant enhancements in tracking metrics. Similarly, in the MultiViewX dataset, the two modules show performance improvement by complementing each other.

### 4.7 Lifting Method Contributions

The TrackTacular [30] proposed various lifting methods and provided results for each. To assess the impact of the two proposed methods, we experimented with different lifting methods on the WildTrack dataset. Our results show significant performance improvements across all lifting methods, as reported in Table 5.

### 4.8 Qualitative Results

The qualitative results are visualizations of the occupancy map, the MVDet network, consistency loss settings, and ground-truth results. (Fig. 4) If you look at the existing MVDet results, you can see blurry points (red boxes). However, when consistency loss is added, you can see that the blurry parts are printed as clear points.

### 4.9 Effectiveness of MPCL and MCFA on MVDet

To validate the generalization of our proposed MPCL and MCFA modules, we incorporated them into MVDet and evaluated the results. The enhanced MVDet achieved superior performance across all metrics compared to the baseline (Table. 6), with significant improvements observed in MODA and Recall. These findings demonstrate that our methods not only improve detection performance but are also effective when applied to networks beyond TrackTacular, highlighting their versatility and impact. (Qualitative Result: (Fig. 4))

## 5 Conclusion

In this work, we proposed a Multi-pedestrian Consistency Loss (MPCL) and a Multi-Camera Feature Aggregation (MCFA) module to address the challenges of the multiview pedestrian detection and tracking tasks. Our approach demonstrates superior performance in preserving the consistency of person features

Table 6: Performance comparison on the WildTrack and MultiViewX datasets, when consistency loss and the multi-camera feature aggregation module are integrated into the MVDet [16] network.

| **Method** | **WildTrack** | | | |
|---|---|---|---|---|
| | **MODA** | **MODP** | **Precision** | **Recall** |
| MVDet | 88.2 | 75.7 | 94.7 | 93.6 |
| + Our method | **91.8** | **76.2** | **95.4** | **96.4** |
| | **MultiViewX** | | | |
| | **MODA** | **MODP** | **Precision** | **Recall** |
| MVDet | 83.9 | 79.6 | 96.8 | 86.7 |
| + Our method | **92.2** | **82.6** | **98.4** | **93.7** |

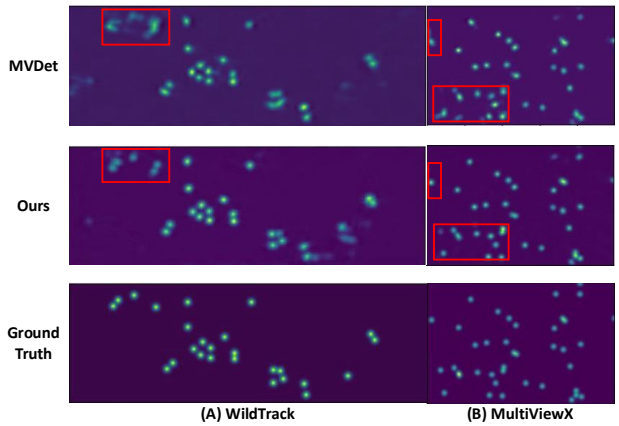


Fig. 4: Visualization of occupancy maps on the WildTrack and MultiViewX datasets. From top to bottom: MVDet [16], Ours, and Ground-truth.

across views, crucial for maintaining robust representations in tasks involving multiple inputs. Additionally, utilizing our multi-camera feature aggregation module, we effectively aggregate features projected from multiple views, leveraging large receptive field convolutions to enhance detection accuracy. Our method achieves state-of-the-art (SOTA) performance on two benchmark datasets, WildTrack and MultiViewX. Specifically, we obtained outstanding results in key detection metrics, achieving SOTA performance in MODA and Recall. In the tracking evaluation, our approach also led to SOTA results in IDF1 and MOTA, further validating the effectiveness of our proposed modules in both detection and tracking tasks.

## Acknowledgements

This work was supported by DeepingSource.

## References


1. Aharon, N., Orfaig, R., Bobrovsky, B.Z.: Bot-sort: Robust associations multi-pedestrian tracking. arXiv preprint arXiv:2206.14651 (2022)
2. Baqué, P., Fleuret, F., Fua, P.: Deep occlusion reasoning for multi-camera multi-target detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 271–279 (2017)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
4. Chavdarova, T., Baqué, P., Bouquet, S., Maksai, A., Jose, C., Bagautdinov, T., Lettry, L., Fua, P., Van Gool, L., Fleuret, F.: Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5030–5039 (2018)
5. Chavdarova, T., Fleuret, F.: Deep multi-camera people detection. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). pp. 848–853. IEEE (2017)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1597–1607 (2020)
7. Cheng, C.C., Qiu, M.X., Chiang, C.K., Lai, S.H.: Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10051–10060 (2023)
8. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
10. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Engilberge, M., Liu, W., Fua, P.: Multi-view tracking using weakly supervised human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1582–1592 (2023)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9729–9738 (2020)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hou, Y., Zheng, L.: Multiview detection with shadow transformer (and view-coherent data augmentation). In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 1673–1682 (2021)
16. Hou, Y., Zheng, L., Gould, S.: Multiview detection with feature perspective transformation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 1–18. Springer (2020)

17. Lee, W.Y., Jovanov, L., Philips, W.: Multi-view target transformation for pedestrian detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 90–99 (2023)
18. Li, S., Ke, L., Danelljan, M., Piccinelli, L., Segu, M., Van Gool, L., Yu, F.: Matching anything by segmenting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18963–18973 (2024)
19. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–18 (2022)
20. Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Hu, Z., Yan, C., Yang, Y.: Improving person re-identification by attribute and identity learning. Pattern Recognition (2019). https://doi.org/https://doi.org/10.1016/j.patcog.2019.06.006
21. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
22. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8844–8854 (2022)
23. Ong, J., Vo, B.T., Vo, B.N., Kim, D.Y., Nordholm, S.: A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(5), 2246–2263 (2020)
24. Qiu, R., Xu, M., Yan, Y., Smith, J.S., Yang, X.: 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In: European Conference on Computer Vision. pp. 695–710. Springer (2022)
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 779–788 (2016)
26. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
27. Song, L., Wu, J., Yang, M., Zhang, Q., Li, Y., Yuan, J.: Stacked homography transformations for multi-view pedestrian detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6049–6057 (2021)
28. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
29. Teepe, T., Wolters, P., Gilg, J., Herzog, F., Rigoll, G.: Earlybird: Early-fusion for multi-view tracking in the bird's eye view. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 102–111 (2024)
30. Teepe, T., Wolters, P., Gilg, J., Herzog, F., Rigoll, G.: Lifting multi-view detection and tracking to the bird's eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 667–676 (2024)
31. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)
32. You, Q., Jiang, H.: Real-time 3d deep multi-camera tracking. arXiv preprint arXiv:2003.11753 (2020)