EarlyBird: Early-Fusion for Multi-View Tracking in the Bird's Eye View

Torben Teepe*   Philipp Wolters   Johannes Gilg   Fabian Herzog   Gerhard Rigoll

Technical University of Munich

Abstract

Multi-view aggregation promises to overcome the occlusion and missed detection challenge in multi-object detection and tracking. Recent approaches in multi-view detection and 3D object detection made a huge performance leap by projecting all views to the ground plane and performing the detection in the Bird's Eye View (BEV). In this paper, we investigate if tracking in the BEV can also bring the next performance breakthrough in Multi-Target Multi-Camera (MTMC) tracking. Most current approaches in multi-view tracking perform the detection and tracking task in each view and use graph-based approaches to perform the association of the pedestrian across each view. This spatial association is already solved by detecting each pedestrian once in the BEV, leaving only the problem of temporal association. For the temporal association, we show how to learn strong Re-Identification (re-ID) features for each detection. The results show that early-fusion in the BEV achieves high accuracy for both detection and tracking. EarlyBird outperforms the state-of-the-art methods and improves the current state-of-the-art on Wildtrack by +4.6 MOTA and +5.6 IDF1. https://github.com/tteepe/EarlyBird

1. Introduction

Detection and tracking of pedestrians has been an essential problem with numerous applications in video surveillance, autonomous vehicles, and sports analysis. Despite the progress on monocular Multiple Object Tracking (MOT) occlusion remains one of the biggest challenges in this research field. Occlusion causes detections to get lost and tracks to get fragmented, thus limiting the detection and tracking quality. However, practical situations like sports analysis require detections in highly cluttered or crowded scenes. Multiple cameras with an overlapping field of view might be available for these cases. Observing a scene from multiple views can help overcome these occlusions since objects hidden in one camera can be visible in another. The challenge then is to aggregate information from multiple
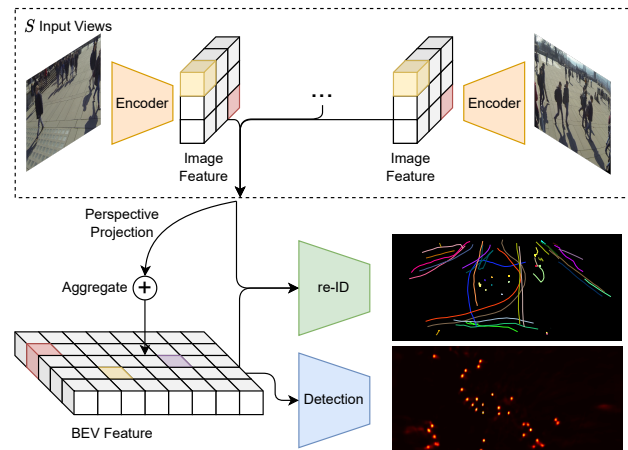

Figure 1. Overview of our approach. All input images are encoded and then perspectively projected to the ground plane. The aggregation reduces the BEV feature where we detect pedestrians and predict a re-ID feature for tracking.

camera views. In early approaches multi-view detection was solved with late fusion methods [46]: First, pedestrians are detected in a single view, then this detection is projected to the 3D space or mostly the ground plane where it is associated with the projections of the other views. More recent approaches [21, 22] utilize an early-fusion strategy that first projects a representation of all views to the common ground plane or *Bird's Eye View* and then perform the detection. These early-fusion detectors [21, 22] increased the detection quality significantly compared to the previous late-fusion approaches. Late-fusion approaches commonly have the advantage that they require less hardware because the processing can be performed independently, and the information projected to 3D is more sparse than the full images. Early-fusion approaches have the advantage that they can be trained end-to-end, while late-fusion usually optimizes the detection and the multi-view association separately. A challenge for the detection in the Bird's Eye View (BEV)-space has been the distortion created by the perspective transformation. Several approaches [21, 27, 39] tried to overcome this problem. We build our approach on [22] but

---

*Correspondence to t.teepe@tum.de

add a BEV-Decoder that is based on a ResNet-18 and gives the decoded features a larger receptive field, allowing the model to aggregate information from the distortion *shadows* to the actual location. We mainly focus on the tracking task, but our model also achieves competitive results in the detection task.

While early-fusion has been shown to be the stronger approach for detection, tracking in multi-view is still performed with the late-fusion approach [8, 20, 40, 41]: first 2D detections are acquired. Secondly, detections of each timestep are associated, and finally, the detections are associated across timesteps. Other approaches [18, 29] switch the order and first associate within one view and later match these tracks across the views. Regardless of the ordering, any stage in this tracking pipeline suffers from inaccuracies introduced by the prior stage, i.e., missed 2D detection later needs to be compensated in the association stage. Our approach combines the first two steps and directly performs the detections in the BEV building upon the latest multi-view detectors [21]. For tracking, we adopt the idea introduced by FairMOT [48] and simultaneously learn a Re-Identification (re-ID) feature for each detection in the BEV-space. This approach allows us to skip the first step of spatial association since our learned detector already solves this problem. The associate in the temporal domain is first performed with appearance-based re-ID features and secondly with a Kalman filter [24] as a motion-based model. We call this architecture EarlyBird. It is an online, end-to-end, trainable tracking architecture that improves the state-of-the-art in tracking by a large margin.

Our contributions are the following:
1) We introduce early-fusion tracking in the Bird's Eye View with a simple but strong re-ID association strategy.
2) We introduced a more robust decoder architecture for the BEV features that improve our tracking results and detections.
3) In our experiments, we qualitatively and quantitatively verify the effectiveness of our method against recent methods and improve the state-of-the-art in tracking on *Wildtrack* by a +4.6 MOTA and +5.6 IDF1.

2. Related Work

Multi-View Object Detection. Using a multiple-camera setup is a prevalent solution to address the difficulties of pedestrian detection with heavy occlusions. Such a setup utilizes synchronized and calibrated cameras observing the same area from different perspectives. The multi-view detection system then integrates these images, all of which have overlapping fields of view, to perform pedestrian detection. Probabilistic modeling of objects [9, 36] was the primary focus before the advancements brought by deep learning. Techniques such as mean-field inference [1, 13]

and conditional random field (CRF) [1, 35] were commonly employed for the aggregation of information from multiple views. However, these techniques often necessitated additional computations or specific designs not inherent in deep learning models. MVDet [22] proposed a convolution-based, end-to-end trainable method that projects encoded image features from each view to the common ground plane, yielding significant improvements and making it the base architecture for all following approaches, including ours. Instead of projecting only the sparse detection from each view to the ground plane, [22] first applies an encoder to the input image and projects all features to the ground plane with perspective transformation. The perspective transformation, which projects image features that depict areas over the ground plane of the actual location in 3D, causes distortions in the ground plane resembling a shadow of the actual object [21]. Other approaches [21, 27, 39] try to overcome these shortcomings of the perspective transformation: [21] uses projection-aware transformers with deformable attention in the BEV-space to aggregate those *shadows* back to the original location. [27] uses regions of interest from the 2D detections and separately projects those to the estimated foot location on the ground plane. Another approach [39] aims to overcome the shortcomings of perspective transformation by using multiple stack homographies at different heights to approximate a complete 3D projection. Instead of focusing on the model side, [32] tried to improve detection on the data side. This approach added additional occlusions with 3D cylindrical objects. This data augmentation makes it harder for the approach to always rely on multiple cameras and thus helps to avoid overfitting.

Our approach builds upon MVDet [22] because it is a solid and straightforward baseline for early-fusion multi-view object detection that we can extend with our tracking approach.

Multi-Target Multi-Camera Tracking. There is much literature on single-camera tracking, and we will discuss one-shot trackers later, but in this section, we discuss the relevant works in Multi-Target Multi-Camera (MTMC) tracking. Most of MTMC trackers assume an overlapping Field of View (FOV) between the cameras. Fleuret *et al*. [13] use the overlapping FOV to model targets into a probabilistic occupancy map (POM) and combine occupancy probabilities with color and motion attributes in the tracking process. As an improvement [2], formulate tracking in POMs as an integer programming problem, and compute the optimal solution by using the k-shortest paths (KSP) algorithm. The problem of MTMC tracking can also be seen as a graph problem. Hypergraphs [20] or multi-commodity network flows [26, 38] are used to model the correspondences across the views and then solved with min-cost [20, 38] or with branch-and-price algorithms [26].

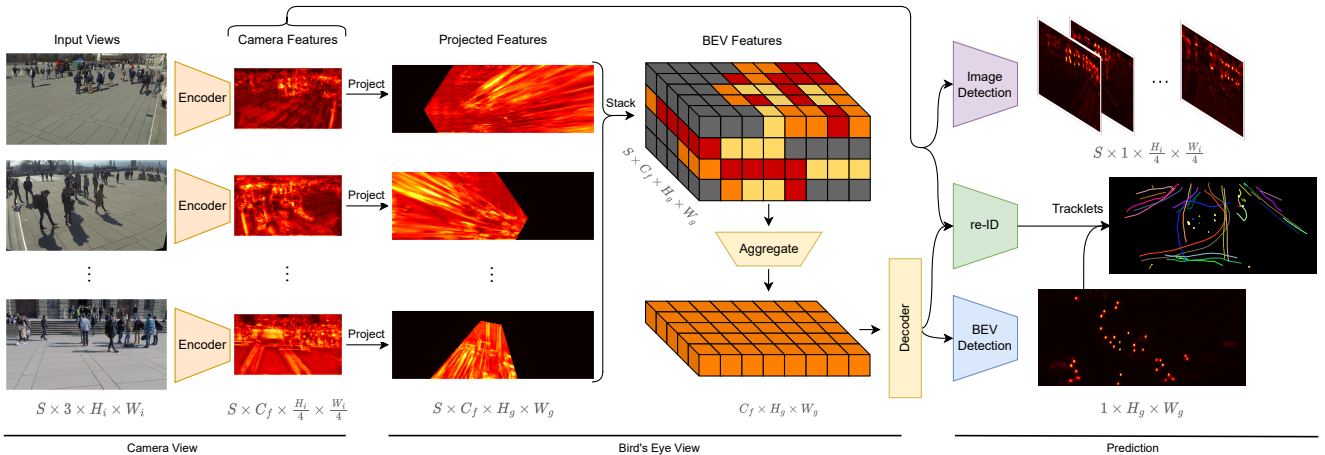In recent years, a two-step approach has become popu-

Figure 2. Overview of our approach. The input view are encoded and the resulting camera features are projected to the ground plane. The projected features are then stacked and aggregated to yield the BEV feature. For the image features the box centers are predicted to guide the occupancy detection in the BEV. Additionally we train a re-ID feature that is guided both by the camera features as well as the BEV features. The detections and their corresponding re-ID features are then used to associate the detections into tracklets.

lar: first generating local tracklets of all the targets within each camera, later matching local tracklets that belong to the same target across all the cameras. For the first step, the generation of local tracklets within a single camera is referred to as the single camera MOT, which has been studied intensively [3, 7, 11, 44, 45, 48, 49]. Due to the impressive progress of object detection techniques, tracking-by-detection [3, 11, 37, 45, 49] has become the mainstream approach for multi-target tracking in recent years. For the second step, various cross-view data association methods have been proposed to match local tracklets across different cameras. Some works [10, 23] use the properties of the epipolar geometry to find correspondences based on location on the ground plane. In addition to ground plane locations [46] adds appearance features as cues for the association. The current state-of-the-art models [8, 29] flip the first two steps: the 2D detections are first projected to the 3D ground plane, and a graph is constructed with re-ID node features. The nodes are then either first assigned spatially and temporally [8] or both assignments happen in the same step [29] using graph neural networks for link prediction. While all current approaches [3, 8, 37, 48, 49] evaluate on detection results to also account for detection inaccuracies, LMGP [29] evaluates on groundtruth bounding boxes and thus can not be compared to any recent works. Our approach differs from all previous work and is more comparable to one-shot trackers covered in the next section. Our approach shares the idea with latest approaches [8, 29] to first associate spatially in our detector and then associate on the ground plane.

One-Shot Tracking. A special case of single-view Multi-

Object Trackers is one-shot trackers. These trackers perform the detection and tracking in one step, thus reducing inference time. They usually have a lower performance compared to two-step trackers. The features predicted can ether be re-ID feature [43, 44, 48] or motion cues [3, 11, 49]. The first example for a re-ID-based approach is Track-RCNN [43] that adds a re-ID feature extraction on top of Mask R-CNN [17] and regresses a bounding box and a re-ID feature for each proposal. Similarly, JDE [44] is build upon YOLOv3 [33], and FairMOT is build upon Center-Net [50]. The advantage of FairMOT compared to the others is that it is anchor-free, meaning detections are not based on bounding boxes but on a single detection point, leading to better separation of the re-ID features. D&T was proposed as a motion-based tracker in [11], which takes input from adjacent frames and predicts inter-frame offsets between bounding boxes. Tracktor [3] directly exploits the bounding box regression head to propagate identities of region proposals and thus removes box association. Unlike other methods, CenterTrack [49] predicts the object center offset on a triplet input: current frame, last frame, and the heatmap of last frame detection. The previous heatmap allows this method to match objects anywhere, even if the boxes overlap. However, motion-based methods only associate objects in adjacent frames without re-initializing lost tracks and thus have difficulty handling occlusions.

In our approach, we thus bring the concept of joint detection and re-ID extraction from FairMOT [48] to MTMC tracking. While training re-ID features for images is well-understood task [19, 43, 44, 48], projecting strong re-ID features to the BEV is what we will investigate in this work.

3. EarlyBird

We provide a comprehensive overview of EarlyBird in Fig. 2. It starts with the input images that are augmented and fed to the encoder network to yield our image features. The image features have the size of the input images downsampled by 4. The image features from all cameras are subsequently projected to the ground plane and stacked into the BEV space. In the following step the BEV space is then reduced in the vertical dimension. The BEV features are finally fed through a decoder network. Both image features and BEV features have separate heads for center and offset detection but share a head for re-ID prediction.

3.1. Encoder

Our approach assumes synchronized RGB input images from S cameras with an input size of $3 \times H_i \times W_i$. We encode the features of the images with ResNet or Swin Transformer networks using three blocks of the network, with each block downsampling the input by 2. Our goal is to only downscale the images by the factor of 4, and thus we upsample and concatenate the output features of each layer until we get an output of $C_f \times H_f \times W_f$ with $H_f = H_i/4$, $W_f = W_i/4$ and $C_f = 128$.

3.2. Projection

The projection is the central part of this approach as it gives a parameter-free link between the image view and the BEV-view. Following [22], we use perspective projection to project the image features to the ground plane. Using the pinhole camera model [15], translation between 3D locations (x, y, z) and 2D image pixel coordinates (u, v) are calculated with:

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \boldsymbol{K} \left[\boldsymbol{R} | \boldsymbol{t} \right] \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}, \quad (1)$$

where s is a real-valued scaling factor, $\boldsymbol{P} = \boldsymbol{K}\left[\boldsymbol{R}|\boldsymbol{t}\right]$ is a 3×4 perspective transformation matrix, $\boldsymbol{K}$ are intrinsic camera matrix and, $[\boldsymbol{R}|\boldsymbol{t}]$ is the 3×4 extrinsic parameter matrix. Eq. (1) describes the ray corresponding to each pixel (u, v) in the 3D world. In our approach, we choose to project all pixels to the ground plane $z = 0$, then the projection can be simplified to:

$$s \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \boldsymbol{P_0} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{14} \\ p_{21} & p_{22} & p_{24} \\ p_{31} & p_{32} & p_{34} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \quad (2)$$

where $\boldsymbol{P_0}$ denotes the 3×3 perspective transformation matrix without the third column from $\boldsymbol{P}$. We apply Eq. (2) to project the features from all S cameras, with their projection $\boldsymbol{P_0^{(s)}}$, to the ground plane grid of a predefined size $[H_g, W_g]$. The size of the ground plane grid depends on the size of the observed and annotated area. Each grid position represents an area of $10\,\mathrm{cm} \times 10\,\mathrm{cm}$, downsampling the annotation grid further by 4 due to memory concerns. All stacked feature maps with C-channels from S cameras give us BEV feature of size $S \times C_f \times H_g \times W_g$.

3.3. Aggregation & Decoder

The goal of the aggregation stage is to combine the features from all S cameras into a single feature, i.e., reduce the S-dimension of the BEV feature map. We concatenate all feature maps along the channel dimension, as in $S \times C_f \times H_g \times W_g \to (S \cdot C_f) \times H_g \times W_g$, yielding a high-dimensional BEV feature map. With two 2D convolutions, we reduce this high-dimensional BEV feature to our desired channel size of $C_g = 128$.

After the aggregation, we feed the BEV feature into a ResNet-18 decoder. The goal of the decoder is to introduce a large receptive field of the ground plane. The distortion introduced by the perspective projections causes pedestrian features to spread out from their actual location on the ground plane. Other approaches [21, 27, 32, 39] identified this distortion as harmful to the detection accuracy and all proposed complex solutions, like deformable transformers [21] or ROI projection [27]. Our decoder offers a simple solution to aggregate location and identification features on the ground plane.

In each layer of the ResNet, the BEV feature is downsampled by 2. We then use a pyramid network architecture to upsample the output of each layer to the size of the previous larger output. Then, both features are concatenated in the channel dimension, and a 2D convolution is applied. The feature pyramid yields a decoded output with the same shape as the input of $C_g \times H_g \times W_g$ but a much higher receptive field for each grid location.

3.4. Heads & Losses

To get the final prediction of the POM, we use prediction heads on our BEV feature map. The detection architecture follows CenterNet [50], and we add a head for center detection that reduces the feature to $1 \times H_g \times W_g$ to yield a heatmap or POM on the ground plane. We add another head for offset prediction that helps predict the location more accurately as it mitigates the quantization error from the ground grid. The offset has an (x, y) component and has the shape $1 \times H_g \times W_g$. Each head is implemented by applying a 3×3 convolution (with $C_g = 128$ channels), followed by an activation layer and a 1×1 convolution to the final target size. The center head is trained with Focal Loss, and the offset head is trained with L1 Loss.

We also add detection heads for image features that predict the center of the 2D bounding boxes and estimated foot location at the bottom-center of the bounding box, helping the image features to have higher activations at the loca-

tion of each pedestrian. Following FairMOT [48], we add an uncertainty term to automatically balance the single-task losses before summing them up.

Re-Identification. The re-ID head aims to generate features that can distinguish individual pedestrians. Ideally, affinity among different pedestrians should be smaller than between the same pedestrian. To archive this, we learn re-ID features through a classification task and as a metric learning task. First, we apply a head that yields the re-ID feature on the ground plane $C_{id,g} \times H_g \times W_g$ with $C_{id} = 64$ and also of the image features $C_{id,f} \times H_f \times W_f$. Afterwards, we extract the feature at the location of the center detection in both planes. We create a class identity distribution with a linear layer that we train with Cross Entropy Loss to the ground truth class identity. As discussed earlier, the perspective transformation introduces strong distortion on the ground plane. Thus, we supervise the re-ID features from the image view. In addition to the Cross-Entropy loss, we apply SupCon Loss [25], which pulls features belonging to the same class identity together while simultaneously pushing apart features of samples from different classes.

3.5. Inference

At inference time, we take the POM predicted by the BEV center head and perform non-maximum suppression (NMS) by a simple 3×3 max pooling operation as in [49]. We then only extract the detections over a certain threshold of 0.4. We also extract the identity embeddings at the estimated pedestrian centers. In the next section, we discuss how we associate the detected boxes over time using the re-ID features.

Online Association. We adopt the hierarchical online data association approach described by MOTDT [7], but instead of boxes, we only track the pedestrians centers seen from the *Bird's Eye View*. Our first step involves initializing a set of tracklets based on the centers detected in the initial timestep. As each subsequent timestep is processed, we connect the centers detected to the existing tracklets using a two-stage matching strategy.

In the first stage, we use a combination of the Kalman Filter [24], and re-ID features to achieve initial tracking results. Specifically, we use the Kalman Filter to anticipate tracklet locations in the next frame and calculate the Mahalanobis distance (D_m) between the anticipated and detected center, similar to the DeepSORT method [45]. We then combine the Mahalanobis distance with the cosine distance computed on re-ID features into a singular distance measure (D) using the formula $D = \lambda D_r + (1 - \lambda)D_m$, where λ is a pre-determined weighting parameter set to 0.98 in our experiments. The Mahalanobis distance is manually set to infinity if it exceeds a certain threshold, which aligns with the JDE protocol [44] and prevents the tracking of trajectories exhibiting implausible motion. We then use the

Hungarian algorithm with a matching threshold $\tau_1 = 0.4$ to conclude the first matching stage.

The second stage involves attempting to match undetected boxes and tracklets based on the center distance of their respective boxes, with an increased matching threshold $\tau_2 = 2.5\,\mathrm{m}$. We continually update the appearance features of the tracklets at each timestep to account for potential variations in appearance. Any unmatched centers are classified as new tracks, and unmatched tracklets are retained for 10 timesteps to facilitate recognition if they reemerge later.

4. Experiments

4.1. Dataset & Metrics

Wildtrack Dataset. Wildtrack [5] is a real-world dataset captured using seven synchronized and calibrated cameras with an overlapping field-of-view of an area of $12\,\mathrm{m} \times 36\,\mathrm{m}$. The movement of the pedestrians is in a public environment and unscripted. Annotations are provided on the ground plane quantized into a 480×1440 grid, resulting in grid cells of $2.5\,\mathrm{cm} \times 2.5\,\mathrm{cm}$. The average number of pedestrians per frame is 20, and 3.74 cameras cover each location. Each camera image is recorded at a resolution of 1080×1920 pixels with a frame rate of $2\,\mathrm{fps}$, covering a total of $35\,\mathrm{min}$.

MultiviewX Dataset. MultiviewX [22] is a synthetic dataset generated in a game engine and is built to be a synthetic copy of the Wildtrack dataset. MultiviewX contains views generated by 6 virtual cameras with overlapping field-of-view. The captured area is with $16\,\mathrm{m} \times 25\,\mathrm{m}$ slightly smaller than the area of the Wildtrack dataset. For annotation, the ground plane is quantized into a grid of size 640×1000, where each grid represents the same $2.5\,\mathrm{cm} \times 2.5\,\mathrm{cm}$ squares. The average number of pedestrians per frame is 40, while 4.41 cameras cover each location. The camera resolution (1080×1920), frame rate ($2\,\mathrm{fps}$), and the length ($400\,\mathrm{frames}$) are equal to Wildtrack.

Detection Metrics. Unlike monocular-view detection systems, which evaluate the predicted bounding boxes, multiview detection systems assess the projected ground plane occupancy map. Thus, the comparison to the ground truth is not calculated with the Intersection over Union (IoU) but with the Euclidean distance as proposed in [5]. Detection is classified as true positive if it is within a distance $r = 0.5\,\mathrm{m}$, which roughly corresponds to the radius of a human body. Following previous works [5, 22], we use Multiple Object Detection Accuracy (MODA) as the primary performance indicator, as it accounts for the normalized missed detections and false positives. Additionally, we report the Multiple Object Detection Precision (MODP), Precision, and Recall.

Tracking Metrics. For tracking the metrics are also calculated in the ground plane. We report the common MOT

		Wildtrack				MultiviewX			
		MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
Two-Stage	RCNN & Cluster [46]	11.3	18.4	68	43	18.7	46.4	63.5	43.9
	DeepMCD [6]	67.8	64.2	85	82	70.0	73.0	85.7	83.3
	Deep-Occlusion [1]	74.1	53.8	95	80	75.2	54.7	97.8	80.2
	MVTT [27]	**94.1**	**81.3**	**97.6**	**96.5**	**95.0**	**92.8**	**99.4**	**95.6**
One-Stage	MVDet [22]	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
	SHOT [39]	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
	3DROM† [32]	91.2	76.9	95.9	95.3	90.0	83.7	97.5	92.4
	MVDeTr [21]	**91.5**	**82.1**	**97.4**	94.0	93.7	91.3	99.5	94.2
	EarlyBird	91.2	81.8	94.9	**96.3**	**94.2**	90.1	98.6	**95.7**

Table 1. Evaluation of the detection performance with the state-of-the-art methods on the Wildtrack and MultiviewX datasets. † 3DROM results are without additional data augmentations.

metrics [4] and identity-aware metrics [34], the threshold for a positive assignment is set to $r = 1\,\mathrm{m}$ to normalize the Multiple Object Tracking Precision (MOTP). The primary metrics under consideration are Multiple Object Tracking Accuracy (MOTA) and IDF1. MOTA takes missed detections, false detections, and identity switches into account. IDF1 measures missed detections, false positives, and identity switches. Additionally, we also report Mostly Tracked (MT) and Mostly Lost (ML). These are reported as a percentage of the total count of unique pedestrians present in the test set.

4.2. Implementation Details

The input size of the images is 720×1280 pixels. For augmentation at train time, we follow [14, 21]: we apply random resizing and cropping on the RGB input, in a scale range of $[0.8, 1.2]$ and adapt the camera intrinsic K accordingly. Additionally, we add some noise to the translation vector t of the camera extrinsic to avoid overfitting the decoder. We train the detector using an Adam optimizer with a one-cycle learning rate scheduler with a maximum learning rate of 10^{-3}. We train for 50 epochs and depending on the size of the encoder, with a batch size of $1 - 2$ but accumulate gradients over multiple batches before updating the weights to have a effective batch size of 16. The encoder and decoder network are initialized with pre-trained weights on *ImageNet-1K*. All experiments are conducted using one RTX 3090 GPU.

4.3. Main Results

Detection. In the tracking-by-detection paradigm, good detections are the basis for good tracking results. While our method does not focus on improving the detection, we still need to be close to the state-of-the-art to achieve competitive results. Tab. 1 compares our methods detection performance to previous methods. We first compare our results to

	Wildtrack				
	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓
KSP-DO [5]	73.2	69.6	61.5	28.7	25.1
KSP-DO-ptrack [5]	78.4	72.2	60.3	42.1	14.6
GLMB-YOLOv3 [30]	74.3	69.7	73.2	79.5	21.6
GLMB-DO [30]	72.5	70.1	63.1	**93.6**	22.8
DMCT [47]	77.8	72.8	79.1	61.0	**4.9**
DMCT Stack [47]	81.9	74.6	78.9	65.9	**4.9**
ReST† [8]	86.7	84.9	84.1	87.8	**4.9**
EarlyBird	**92.3**	**89.5**	**86.6**	78.0	**4.9**
	MultiviewX				
	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓
EarlyBird	**82.4**	**88.4**	**86.2**	**82.9**	**1.3**

Table 2. Evaluation of tracking results on the Wildtrack and MultiviewX. † ReST originally reported the tracking metrics on view-based tracking instead of tracking in the projected view. The results shown are re-computed by us.

the baseline: MVDet [22], as we base our approach on it. The results show that our decoder architecture and augmentation changes improved MVDet. Other detection-focused methods [21, 27, 39] also extend MVDet and achieve comparative results on Wildtrack, but our approach has competitive results on MultiviewX. The current state-of-the-art MVTT [27] is a two-stage detection approach and could still be added to our single-state approach to further improve the results.

Tracking. In Tab. 2 we compare our method to state-of-the-art approaches. Our approach outperforms all current approaches by a big margin. Compared to the current best-performing method ReST [8], we improve the IDF1 by 5.6, and the MOTA by 4.6 percent points. All other meth-

ods [5, 8, 29, 30, 47] start from 2D detections, and to compare tracking methods, you also need to take the detection quality into account. For most of these approaches, we cannot directly compare the detection quality, but ReST [8] uses detection from MVDeTr [21], which performs very close to our detector (*cf*. Tab. 1). ReST and EarlyBird follow a similar approach: associate spatially on the ground plane, then associate temporally. However, ReST only projects detections in 2D to the ground plane and associates them with a graph solver. In contrast, we project the complete input image feature space to the ground plane and associate it with the decoder. Our results show that with similar detection quality, our approach outperforms ReST, which shows the advantage of our early-fusion approach compared to graph-based late-fusion.

4.4. Ablations Studies

Influence of Method Components. Next, we ablate each component introduced in our method compared to the baseline as shown in Tab. 3. The baseline consists of MVDet [22] with minimal additions to perform tracking, namely a ReID head added to the BEV-space. The baseline results suffer from strong overfitting and the added augmentation in the next step mostly aleviates this. We augment the input images with scaling and cropping to avoid overfitting in the encoder and add transitive noise to the projection to the ground plane to avoid overfitting in the prediction heads. With better augmentation, we introduced our larger decoder based on a ResNet-18 with a feature pyramid. These additions gave us one of the most robust detection results, but as tracking is our primary focus, we added the view center and re-ID loss. These two losses are applied to the image features and should help guide these features. While the 2D center detection alone decreased our detection performance, using it with the re-ID loss gave us the final SOTA results.

Influence of the Encoder Network. The encoder extracts the features of the RGB image that are projected to the ground plane. While all of the similar detection-based approaches [21, 22, 27, 32] use a ResNet-18 encoder, these approaches do not need to encode the identity feature. Thus, we try our approach with larger encoders and transformer-based encoders, see Tab. 4. The ablation shows that ResNet-18 has the best performance. ResNet-50 may be slightly better in the detection and tracking performance, but the smaller ResNet-18 outperforms it in the main metrics MODA and MOTA and has competitive scores for IDF1. We thus use ResNet-18 to report all other results.

4.5. Qualitative Results

In Fig. 3, we plot the output of our model on the test set of Wildtrack. In Fig. 3a, we compare the prediction of the POM to the ground truth as a heatmap at a single timestep. Each point in the ground truth represents a pedestrian in the

	Detection		Tracking		
	MODA	MODP	IDF1	MOTA	MOTP
Baseline	77.8	78.9	71.3	72.6	80.9
+ Augmentation	89.5	81.7	84.5	87.4	83.0
+ Decoder	**91.3**	**82.2**	91.1	89.1	**86.9**
+ View Center Loss	91.0	82.1	90.0	89.1	84.0
+ View re-ID Loss	91.2	81.8	**92.3**	**89.5**	86.6

Table 3. Ablation of the components introduced by our approach compared to the baseline method.

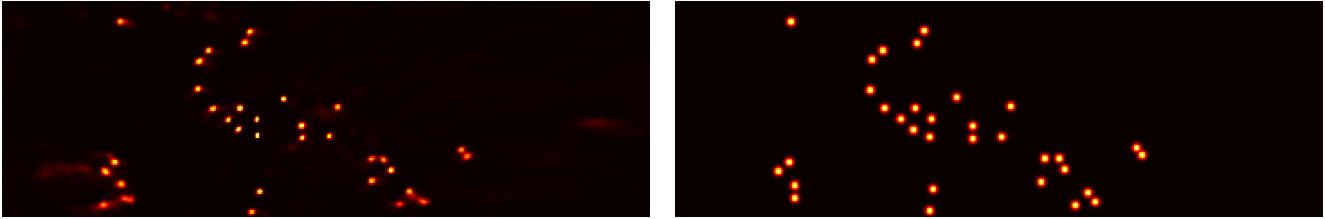|	Detection		Tracking		
	MODA	MODP	IDF1	MOTA	MOTP
ResNet-18	**91.2**	81.8	92.3	**89.5**	86.6
ResNet-50	89.6	**82.3**	**92.6**	88.8	86.5
Swin-T	89.5	81.3	92.0	87.3	**87.9**

Table 4. Ablation of different encoders on detection and tracking results of the Wildtrack dataset.

BEV. The scene's center is crowded with pedestrians, and the prediction in this high-overlap area is almost perfect. The further the pedestrians are on a side, the less accurate the prediction is. This inaccuracy could be due to the increased distortion further from the cameras and less overlap of the views in the border regions.
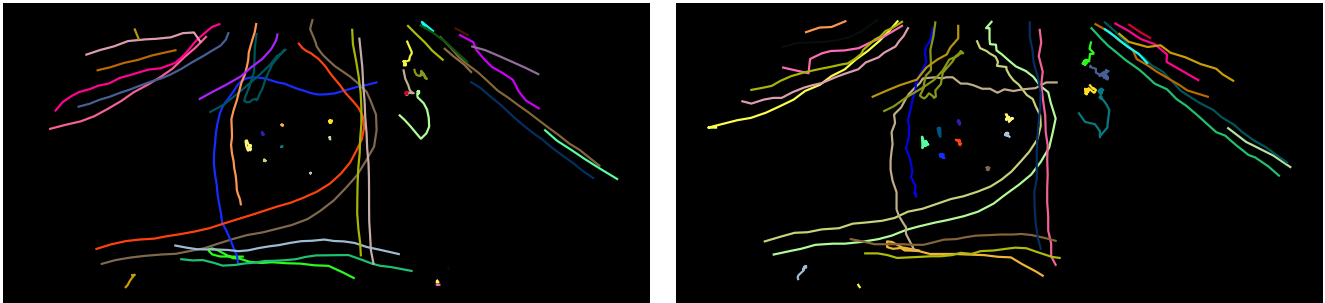
Fig. 3b shows similar results for tracking. We show all tracks on the ground plane of the full test set of Wildtrack, where each color and line represents the path one identity takes. The tracks in the center of the scene are predicted almost perfectly. However, the top-left and top-right tracks are segmented or switched tracks. This inaccuracy could be due to less accurate or missing detections and identity features outside the ground plane.

5. Discussion

Limitations. The first limitation of our approach is the requirement for high-quality 3D annotations and camera calibrations. While this is easy to archive with synthetic data, it is costly for real-world data. Therefore, we could not evaluate some older datasets (CAMPUS [46], PETS09 [12]), where most of the late-fusion models can work with only 2D annotations and ground plane homography. Furthermore, our approach requires synchronized cameras. Since we lift all cameras to the same 3D space, the time differences should be minimal so that moving objects do not project to different locations in 3D. Late-fusion MTMC tracking methods can account for more drift in the temporal domain. Our approach also has higher hardware requirements. While late-fusion approaches may process each

(a) Comparison of the detection result on Wildtrack as a heat map. Each point in the ground truth represents a pedestrian in the BEV.

(b) Comparisons of the tracking results on Wildtrack. Each colored line represents a path taken by one tracked pedestrian as seen from the BEV.

Figure 3. Qualitative detection (a) and tracking (b) results of our approach (left) compared to the ground truth (right).

camera decentralized and fuse the information centrally, our approach processes all camera images simultaneously. It thus requires more memory and computational resources on a single machine.

Ethical Impact. Tracking methods always have the risk of being used for illegal surveillance. Methods that focus on pedestrian tracking must face this criticism, especially. The dataset Wildtrack [5] has been criticized [16] for the missing consent of the recorded persons. Unfortunately, a good comparison to the state-of-the-art is only possible with this dataset, even though a synthetic replication with MultiviewX [22] is now available. We are the first to evaluate tracking on this synthetic dataset to lower the ethical implications of tracking.

Future Work. For the detection part, the biggest challenge of our and other current approaches [21, 27, 39] is the distortion caused by the projective transformation. Other methods that lift from 2D to 3D space could be explored for multi-view detection, i.e., Simple-BEV [14], Lift-Splat-Shot [31], or BEVFormer [28]. Most current approaches only use the current frame for detection. Using more temporal frames could improve the detection performance [28]. Using more temporal context could also improve the tracking quality, and approaches like CenterTrack [49] could be used to track via motion cues. The pedestrian datasets used in this work have about 400 timestamps, which is relatively small by modern computer vision standards, and the detection and tracking accuracy is saturating. The need for larger datasets is apparent, and datasets with similar MTMC prob-

lems for traffic surveillance [18, 42] could bridge this gap.

6. Conclusion

This paper shows that the EarlyBird catches the worm through multi-view highly accurate tracking. Early-fusion of all views and tracking in the bird's eye view considerably improves MTMC tracking. We adapt one-shot tracking to multi-view tracking to propose an online, anchor-free tracker. We propose ways to efficiently train re-ID features in BEV and ablate each of our tracking improvements.

We expect EarlyBird to inspire feature work in early-fusion multi-view tracking and believe that EarlyBird, together with our suggested future work, makes significant progress towards tackling multi-view tracking problems.

References

[1] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *ICCV*, pages 271–279, 2017. 2, 6

[2] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *IEEE TPAMI*, 33(9):1806–1819, 2011. 2

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *CVPR*, pages 941–951, 2019. 3

[4] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 6

[5] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *CVPR*, pages 5030–5039, 2018. 5, 6, 7, 8

[6] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 848–853. IEEE, 2017. 6

[7] Long Chen, Haizhou Ai, Zijie Zhuang, and Chong Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *ICME*, pages 1–6. IEEE, 2018. 3, 5

[8] Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, and Shang-Hong Lai. Rest: A reconfigurable spatial-temporal graph model for multi-camera multi-object tracking. *arXiv preprint arXiv:2308.13229*, 2023. 2, 3, 6, 7

[9] Adam Coates and Andrew Y Ng. Multi-camera object detection for robotics. In *2010 IEEE International Conference on Robotics and Automation*, pages 412–419. IEEE, 2010. 2

[10] Ran Eshel and Yael Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *CVPR*, pages 1–8. IEEE, 2008. 3

[11] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, pages 3038–3046, 2017. 3

[12] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–6. IEEE, 2009. 7

[13] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE TPAMI*, 30(2):267–282, 2007. 2

[14] Adam W. Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-BEV: What really matters for multi-sensor bev perception? In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 6, 8

[15] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[16] Jules. Harvey, Adam. LaPlace. Exposing.ai, 2021. 8

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 3

[18] Fabian Herzog, Junpeng Chen, Torben Teepe, Johannes Gilg, Stefan Hörmann, and Gerhard Rigoll. Synthehicle: Multi-vehicle multi-camera tracking in virtual cities. In *WACV Worksh.*, pages 1–11, January 2023. 2, 8

[19] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification. In *ICIP*, pages 1129–1133. IEEE, 2021. 3

[20] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *CVPR*, pages 3650–3657, 2013. 2

[21] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *ACM MM*, 2021. 1, 2, 4, 6, 7, 8

[22] Yunzhong Hou, Liang Zheng, and Stephen Gould. Multiview detection with feature perspective transformation. In *ECCV*, 2020. 1, 2, 4, 5, 6, 7, 8

[23] Weiming Hu, Min Hu, Xue Zhou, Tieniu Tan, Jianguang Lou, and Steve Maybank. Principal axis-based correspondence between multiple cameras for people tracking. *IEEE TPAMI*, 28(4):663–671, 2006. 3

[24] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960. 2, 5

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 5

[26] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Branch-and-price global optimization for multi-view multi-target tracking. In *CVPR*, pages 1987–1994. IEEE, 2012. 2

[27] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Multiview target transformation for pedestrian detection. In *WACV Worksh.*, pages 90–99, 2023. 1, 2, 4, 6, 7, 8

[28] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18, 2022. 8

[29] Duy MH Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, and Paul Swoboda. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In *CVPR*, pages 8866–8875, 2022. 2, 3, 7

[30] Jonah Ong, Ba-Tuong Vo, Ba-Ngu Vo, Du Yong Kim, and Sven Nordholm. A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. *IEEE TPAMI*, 44(5):2246–2263, 2020. 6, 7

[31] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 8

[32] Rui Qiu, Ming Xu, Yuyao Yan, Jeremy S Smith, and Xi Yang. 3d random occlusion and multi-layer projection for deep multi-camera pedestrian localization. In *ECCV*, pages 695–710. Springer, 2022. 2, 4, 6, 7

[33] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[34] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016. 6

[35] Gemma Roig, Xavier Boix, Horesh Ben Shitrit, and Pascal Fua. Conditional random fields for multi-camera object detection. In *ICCV*, pages 563–570, 2011. 2

[36] Aswin C Sankaranarayanan, Ashok Veeraraghavan, and Rama Chellappa. Object detection, tracking and recognition for multiple smart cameras. *Proceedings of the IEEE*, 96(10):1606–1624, 2008. 2

[37] Jenny Seidenschwarz, Guillem Brasó, Víctor Castro Serrano, Ismail Elezi, and Laura Leal-Taixé. Simple cues lead to a strong multi-object tracker. In *CVPR*, pages 13813–13823, 2023. 3

[38] Horesh Ben Shitrit, Jérôme Berclaz, François Fleuret, and Pascal Fua. Multi-commodity network flow for tracking multiple people. *IEEE TPAMI*, 36(8):1614–1627, 2013. 2

[39] Liangchen Song, Jialian Wu, Ming Yang, Qian Zhang, Yuan Li, and Junsong Yuan. Stacked homography transformations for multi-view pedestrian detection. In *CVPR*, pages 6049–6057, 2021. 1, 2, 4, 6, 8

[40] Andreas Specker and Jürgen Beyerer. Reidtrack: Reid-only multi-target multi-camera tracking. In *WACV*, pages 5441–5451, 2023. 2

[41] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *CVPR*, pages 4173–4182, 2021. 2

[42] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR*, pages 8797–8806, 2019. 8

[43] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019. 3

[44] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *ECCV*, pages 107–122. Springer, 2020. 3, 5

[45] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, pages 3645–3649. IEEE, 2017. 3, 5

[46] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *CVPR*, pages 4256–4265, 2016. 1, 3, 6, 7

[47] Quanzeng You and Hao Jiang. Real-time 3d deep multi-camera tracking. *arXiv preprint arXiv:2003.11753*, 2020. 6, 7

[48] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129:3069–3087, 2021. 2, 3, 5

[49] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, pages 474–490. Springer, 2020. 3, 5, 8

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 3, 4