Lifting Multi-View Detection and Tracking to the Bird's Eye View

Torben Teepe Philipp Wolters Johannes Gilg Fabian Herzog Gerhard Rigoll
Technical University of Munich

Abstract

Taking advantage of multi-view aggregation presents a promising solution to tackle challenges such as occlusion and missed detection in multi-object tracking and detection. Recent advancements in multi-view detection and 3D object recognition have significantly improved performance by strategically projecting all views onto the ground plane and conducting detection analysis from a Bird's Eye View (BEV). In this paper, we compare modern lifting methods, both parameter-free and parameterized, to multi-view aggregation. Additionally, we present an architecture that aggregates the features of multiple times steps to learn robust detection and combines appearance- and motion-based cues for tracking. Most current tracking approaches either focus on pedestrians or vehicles. In our work, we combine both branches and add new challenges to multi-view detection with cross-scene setups. Our method generalizes to three public datasets across two domains: (1) pedestrian: Wildtrack and MultiviewX, and (2) roadside perception: Synthehicle, achieving state-of-the-art performance in detection and tracking. *https://github.com/tteepe/TrackTacular*.

1. Introduction

Multi-Target Multi-Camera (MTMC) tracking has been a niche topic within the tracking community compared to the more popular Multiple Object Tracking (MOT) task. Even though using multiple cameras to overcome the challenge of occlusion and missed detections was already introduced in one of the first modern tracking datasets PETS2009 [15]. However, in recent years, the MTMC task gained more attention with more cameras beginning to be deployed in the wild and the availability of more MTMC datasets [10, 18, 24, 27, 43]. In recent years, new approaches were either designed for pedestrian tracking [7, 27, 43] or for tracking vehicles [24]. In this paper, we want to unify both branches and introduce an approach that can generalize to both domains and outperforms the state-of-the-art on four public datasets: two classical pedestrian datasets: Wildtrack [7] and MultiviewX [27] and one vehicle datasets: Synthehicle [24]. The commonly used pedestrian datasets, Wildtrack and MultiviewX, have a few shortcomings for modern computer vision research as they only consist of one scene, and they have a first 90% of frames train and last 10% of frames test split, which is prone to overfitting. Thus, we evaluate a new challenging dataset. Synthehicle [24] is a synthetic roadside dataset covering multiple intersection scenarios for vehicle tracking. This dataset requires approaches to generalize over different unseen scenes.

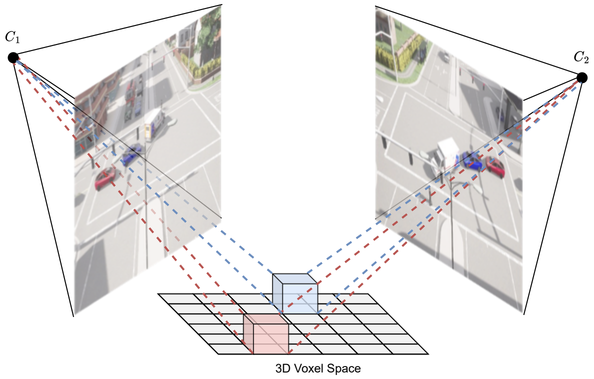


Figure 1. **Lifting Methods.** We compare three methods that lift the pixel information to 3D voxel Bird's Eye View (BEV) space for detection and tracking.

Compared to the single-camera task, the MTMC task is more challenging as it requires two association steps: (1) an association between cameras and (2) an association between detections. On the other hand, multi-camera approaches allow for a much stronger 3D perception as the 3D position of the objects can be triangulated from multiple cameras. Traditionally MTMC approaches [24, 25, 34] start with 2D bounding box detection and then create camera tracklets that are associated to form a global track. Other approaches used 2D detection to predict a 3D location and perform the association in 3D [9]. More recently, approaches started skipping the 2D detection step; they project the features to the BEV and perform both detection and tracking in 3D [44]. These works show that fusing the data earlier in 3D yields better results than late-fusion approaches. We want to apply this early-fusion mindset and propose a new architecture that uses more advanced methods to project the features in a 3D space.

The projection of camera features has been studied extensively in the perception models [20, 32, 36] for autonomous vehicles and has become an essential part of the multi-sensor perception system. While other sensors like Lidar and Radar have depth information, the camera image is a projection of the 3D world onto a 2D plane, with no trivial way to reverse this projection. Thus, lifting methods have been developed to recover 3D information. The projection is also essential to fuse camera data with other sensors [50].

Nevertheless, there are critical differences between vehicle perception and multi-view detection: (1) The cameras in our approach have a larger overlapping area and thus can aggregate information from multiple cameras. (2) Our cameras are mostly static; thus, we can exploit the scene's geometry (3) The observed area is much larger, and the cameras are further away from the targets. However, the overlap gives us a key advantage: We can triangulate the 3D position of the targets using an approximation to *good ol'* epipolar geometry [21].

Tracking is also part of autonomous perception; the challenge is more manageable than tracking with static cameras, as the targets are within a small surrounding area and are only observed shortly. Thus, most methods [46, 52] use simple distance-based association methods. In our approach, we face more significant scenes with more targets that need to be tracked over extended periods. Staying with the early-fusion approach, we propose a novel association method that uses the history BEV information to predict the location of each detection in the previous frame. It combines the advantages of appearance-based and motion-based association and learns to combine both cues.

In summary, our contributions are as follows:

- We combine our novel tracking strategy with three existing lifting methods and extend them for views with strong overlap to show state-of-the-art detection and tracking results on three public datasets.
- We propose a novel learned association method that combines the advantages of appearance-based and motion-based association and outperforms previous methods in tracking.
- We set a new strong baseline on more challenging datasets for MTMC with a standard evaluation protocol to initiate further and more comparable research in this field.

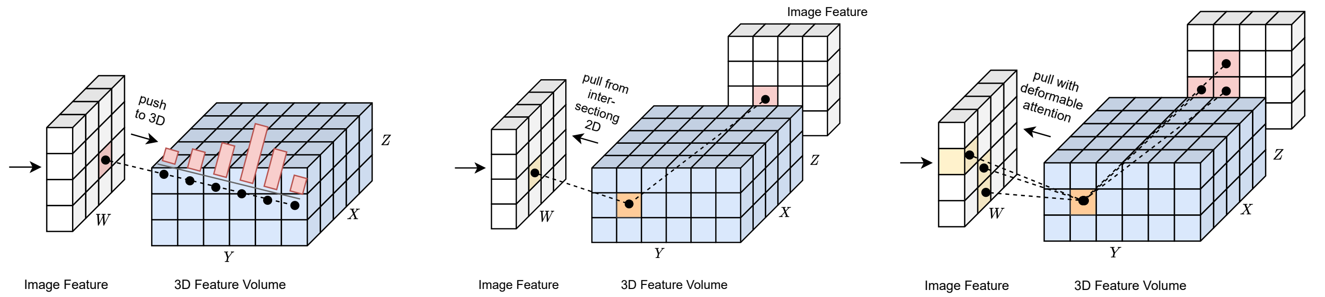
2. Related Work

Multi-View Object Detection. Multi-camera systems are widely used to tackle the challenge of pedestrian detection in highly occluded environments. Such systems comprise synchronized and calibrated cameras that capture a common area from various angles. The multi-view detection system then processes the overlapping views to detect pedestrians. MVDet [27] introduced an approach that uses convolutions to train models end-to-end, projecting encoded image features from each perspective onto a shared ground plane. This process led to noticeable improvements and has been the basis for subsequent methods, including ours. Projections are not just limited to the sparse detections from each viewpoint. The method introduced by [27] encodes the input image and projects all features onto the ground plane using a perspective transformation. This mapping results in distortions on the ground plane like a shadow of the actual object [26]. To overcome the limitations of perspective transformation, several other methods [26, 31, 42] have been proposed. One approach [26] uses projection-aware transformers with deformable attention in the BEV-space to move the ”shadows” back to their original location. Another method [31] uses regions of interest from 2D detections and separately projects these to the estimated foot position on the ground plane. A third approach [42] uses multiple stacked homographies at various heights to better approximate a complete 3D projection. Shifting the focus from model improvement, [37] attempted to enhance detection by improving the data. This approach added more occlusions by introducing 3D cylindrical objects. Such a data augmentation reduces the dependence on multiple cameras, helping to avoid overfitting.

Overall, our approach follows the tracks of MVDet [27], and we include their projection method as our baseline. However, among other improvements, we will also explore other projection methods explained in the next paragraph to improve the detection performance.

3D Perception Systems. Multi-sensor perception systems are mainly developed for autonomous vehicles to fuse the data from different sensors. With the enormous interest in autonomous driving, this area is progressing rapidly. For this section, we want to focus on approaches that focus on camera lifting.

The first and simplest unprojection are homography-based methods [26, 27, 31, 42]. They assume a flat ground plane and use a homography matrix to project the image features to the ground plane. While this method is parameter-free, it is less accurate for objects above the ground plane and causes shadow-like artifacts for objects far away from the camera [26]. Simple-BEV [20] proposed another parameter-free projection method: it defines a 3D volume of coordinates over the BEV plane, projects these coordinates into all images, and averages the features sampled from the projected locations [19]. Compared to homography-based projection, this method *pulls* information from the image to a 3D location instead of *pushing* information from the image to the world. Depth-based approaches [36] employ a monocular depth estimator that es-



(a) In the depth splatting approach [36] the 2D feature are *pushed* to 3D with depth prediction, filling voxels that intersect with its ray.

(b) The bilinear sampling method introduced in Simple-BEV [20] each 3D Voxel *pulls* feature from the 2D map by projection and sampling.

(c) The lifting method introduced by BEVFormer [32] uses deformable attention to aggregate multiple image features with a learnable offset.

Figure 2. **Lifiting Methods.** The three lifting methods we compare in this paper. The bilinear sampling method (b) simplifies the depth splat approach (a) without explicitly predicting the depth. Our method extends the bilinear sampling to only project image features if they intersect in the 3D volume. Thus, our method approximates the triangulation at voxel granularity.

timates a per-pixel depth to project the pixels into the 3D space. This method is very effective as it does not require depth information to be explicitly available and can also deduct it from the scene. Another way of lifting the image features is BEVFormer [32]. It is similar to Simple-BEV [20] as it aggregates from all images to a 3D location. However, it uses a deformable attention mechanism to learn the aggregation weights. This method is more flexible than Simple-BEV [20] as it can learn to focus on specific objects, but it comes with a much higher computational cost. In our work, we will compare these different lifting strategies and extend them to explicitly enforce the triangulation of features.

Multi-Target Multi-Camera Tracking. There is a wealth of research on single-camera tracking, which will be discussed later. However, we concentrate on MTMC tracking in this section. Most MTMC trackers base their models on the assumption of an overlapping Field of View (FOV) among cameras. The method by Fleuret et al. [16] makes use of this overlapping FOV to represent targets within a probabilistic occupancy map (POM), combining color and motion features with occupancy probabilities during the tracking process. [2] enhanced this approach by framing tracking within POMs as an integer programming problem, solving it optimally using the k-shortest paths (KSP) algorithm.

MTMC tracking can also be interpreted as a graph problem. Hypergraphs [25] or multi-commodity network flows [30, 41] are used to establish correspondences across views, which are then resolved using min-cost [25, 41] or branch-and-price algorithms [30].

Recently, a two-stage approach has gained popularity: it begins with generating local tracklets for all targets within each camera, followed by matching local tracklets corresponding to the same target across all cameras. The task of generating local tracklets within a single camera, known as single camera MOT, has been extensively studied [3, 8, 14, 47, 49, 54, 56]. With the advancement in object detection techniques, tracking-by-detection [3, 14, 40, 49, 56] is now the favored method for multi-target tracking. For the second step, several strategies for cross-view data association have been proposed to match local tracklets across different cameras. Some studies [13, 28] use the principles of epipolar geometry to find correspondences based on location on the ground plane. Besides ground plane locations, [51] incorporates appearance features as cues for the association. Cutting-edge models [9, 34] have inverted the initial two steps: the 2D detections are first projected onto the 3D ground plane, and a graph is created using Re-Identification (re-ID) node features. These nodes are either assigned spatially and temporally at the outset [9] or both assignments are made simultaneously [34] utilizing graph neural networks for link prediction. While all current methods [3, 9, 40, 54, 56] are evaluated based on detection results to account for inaccuracies in detection, LMGP [34] uses ground truth bounding boxes and thus cannot be compared to recent studies.

Our approach stands apart from all previous work and aligns more closely with one-shot trackers, which are discussed in the subsequent section. Similar to the latest methods [9, 34], our approach establishes spatial associations within our detector, then proceeds to associate on the ground plane.

One-Shot Tracking. A subset of single-view Multi-Object Trackers includes one-shot trackers. These systems conduct detection and tracking in a single step, saving computation time; however, they generally perform worse than two-step trackers. The features predicted can either be re-ID features [45, 47, 54] or motion cues [3, 14, 56].

Track-RCNN [45] is the first example of a re-ID-based approach. It adds re-ID feature extraction to Mask R-CNN [22], generating a bounding box and a corresponding re-ID

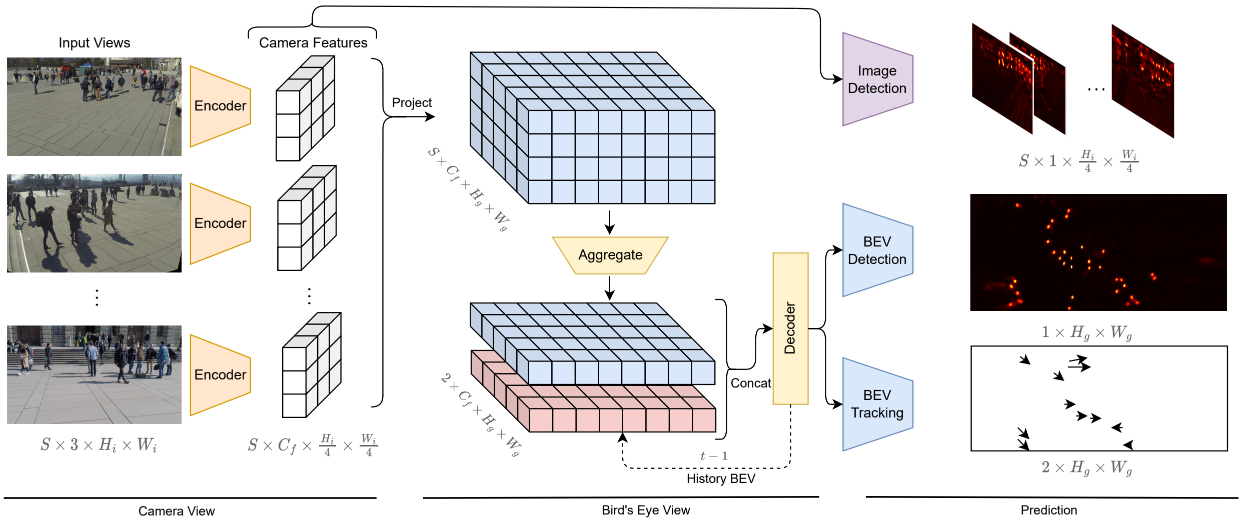

Input Views
Camera Features
Encoder
Encoder
Encoder
Project
Image Detection
$S \times 1 \times \frac{H_i}{4} \times \frac{W_i}{4}$
$S \times C_f \times H_g \times W_g$
Aggregate
Decoder
BEV Detection
BEV Tracking
$1 \times H_g \times W_g$
$2 \times C_f \times H_g \times W_g$
Concat
$t-1$
History BEV
$2 \times H_g \times W_g$
$S \times 3 \times H_i \times W_i$
$S \times C_f \times \frac{H_i}{4} \times \frac{W_i}{4}$
Camera View
Bird's Eye View
Prediction

Figure 3. **Overview of Our Approach.** The input views are encoded, and the resulting camera features are projected using one of three lifting methods. After aggregation, the BEV feature is concatenated with the feature of the previous step. With the decoded BEV feature, we predict the locations and offset to the location in the previous step. Additionally, we guide the architecture by predicting the object centers in the image features.

feature for each proposal. Similarly, JDE [47] is built upon YOLOv3 s[38], and FairMOT is based on CenterNet [55]. The key advantage of FairMOT, compared to the others, is its anchor-free design, meaning that detections do not rely on a fixed set of anchor bounding boxes but on a singular detection point, enhancing the separation of re-ID features.

Motion-based trackers such as D&T [14] take input from adjacent frames, predicting inter-frame offsets between bounding boxes. Tracktor [3] exploits the bounding box regression head to associate identities, thus removing box association. In contrast, CenterTrack [56] predicts the object center offset using a triplet input, consisting of the current frame, the previous frame, and the heatmap of the last frame detection. This prior heatmap allows objects to be matched anywhere, even in the case of overlapping boxes. However, these motion-based methods, which only associate objects in adjacent frames without re-initializing lost tracks, can struggle in managing occlusions.

In our approach, we propose to learn the association between detections in the previous and current time steps, as CenterTrack [56] and D&T [14] proposed. However, we apply this idea at BEV feature level. This approach combines the advantages of appearance-based and motion-based association and learns to combine both cues. Combined with a Kalman Filter [29], we can also re-initialize lost tracks.

3. Methodology

We provide an architecture overview in Fig. 3. It starts with the S RGB input images ($S \times 3 \times H_i \times W_i$) that are augmented and fed to the encoder network to yield our downsampled image features ($S \times C_f \times \frac{H_i}{4} \times \frac{W_i}{4}$). With different projection methods, features are then projected to a common BEV space ($S \times C_f \times H_g \times W_g$). In the following step, the BEV space is then reduced in the vertical dimension ($S \times H_g \times W_g$). The feature of the previous time step is subsequently concatenated to the current BEV feature ($2 \times S \times C_f \times H_g \times W_g$). The BEV features are finally fed through a decoder network.

3.1. Lifting Methods

The projection is central to this approach as it provides the link between the image view and the BEV-view.

Perspective Transformation. This method is the simplest lifting method as it does not model height information. Following [27], we employ perspective projection to transfer the image features onto the ground plane. The pinhole camera model [21] uses a 3×4 transformation matrix $\boldsymbol{P} = \boldsymbol{K}\,[\boldsymbol{R}|\boldsymbol{t}]$ that maps 3D locations (x, y, z) to 2D image pixel coordinates (u, v). We choose to project all pixels to the ground plane at $z = 0$, which simplifies the projection to:

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

where s is a real-valued scaling factor and $\boldsymbol{P_0}$ denotes the 3×3 perspective transformation matrix without the third column from $\boldsymbol{P}$. Features from all S cameras are projected to the ground plane using this equation, with each having its unique projection matrix $\boldsymbol{P_0^{(s)}}$.

Depth Splatting. The Depth Splat method [36] is based on monocular depth estimation. The idea is to simulate a point cloud from a camera image. Each image pixel (u, v) is associated with a discrete depth $d \in D = \{d_0 + \Delta, ..., d_0 + |D|\Delta\}$. This depth distribution is predicted as part of the image features, making this a parameterized lifting method. Unprojecting the image feature channels along the predicted depth yields our point cloud of $(D \times C_f \times \frac{H_i}{4} \times \frac{W_i}{4})$ in each camera frustum. The point clouds are then wrapped into a common voxel space, weighting the channel information with the probably of the discrete depth (*cf*. Fig. 2a). This method is considered to *push* its information from 2D to 3D.

Bilinear Sampling. The idea of Simple-BEV [20] is to simplify the method of depth projection without explicitly predicting the depth. Each ray would fill its information into all voxels that intersect with it. However, the projection is turned around for this method: the 3D voxels *pull* the information from the 2D image. This pulling is done by projecting the 3D voxel to the image plane, determining if the projected point is inside the image, and later sampling the image features sub-pixel accurate to the voxel. We can project all eight vertices of a voxel to all image planes with

$$d \begin{pmatrix} u_n \\ v_n \\ 1 \end{pmatrix} = \boldsymbol{K} \, [\boldsymbol{R}|\boldsymbol{t}] \begin{pmatrix} x_n \\ y_n \\ z_n \\ 1 \end{pmatrix} = \boldsymbol{P}^{(s)} \begin{pmatrix} x_n \\ y_n \\ z_n \\ 1 \end{pmatrix}, \qquad (2)$$

and sample the image feature from $[\min(u_n), \min(v_n), \max(u_n), \max(u_n)]$. The features from all S cameras are then averaged for each voxel. The advantage is that every voxel will receive a feature, while in splatting methods, some voxels further away from the cameras might not be filled at all. This property makes this method more robust in long-range perception. Even though the bilinear sampling method [19, 20] was not designed for multi-camera perception, it reveals a unique property in the overlapping areas: it approximates a triangulation of image points at voxel granularity. This triangulation is illustrated in Fig. 2c. While for data with complete overlap, i.e., every voxel is at least seen by two cameras, the bilinear sampling method is equivalent to the triangulation method.

Deformable Attention. The lifting method introduced in BEVFormer [32] uses an approach that uses each voxel as a query and projects the 3D reference points back to the 2D image views with the Eq. (2). The 2D reference points for each query and features around those image feature locations are sampled. Finally, the features are aggregated as a weighted sum as the output of spatial cross-attention. The approach is similar to bilinear sampling, but the aggregation uses surrounding features of the projected location and aggregates the BEV features with attention.

Instead of the temporal self-attention introduced in BEVFormer, we use the same resource-efficient temporal aggregation introduced in the next section for all lifting methods.

3.2. Temporal Aggregation

The core of tracking is to aggregate temporal information. In our architecture, we want to fuse these features early instead of at the detection level. Temporal information can also improve the detection, as detections can not disappear between time steps. The availability of the previous features enables the architecture to learn the motion of each detection. Trackers are usually divided into appearance-based and motion-based; however, our feature contains both types of information, and our architecture can fuse both cues at the feature level. We implement the temporal aggregation in a *late-to-early* fusion [23] manner: the decoded BEV feature of the previous timestep is concatenated to the current, undecoded feature (*cf*. Fig. 3).

3.3. Detection & Tracking Heads

The general head architectures follow CenterNet [55], and our main detection branch predicts a heatmap or POM on the ground plane. We add another head for offset prediction (x, y) that helps predict the location more accurately as it mitigates the quantization error from the voxel grid. We train the center head with Focal Loss [33], and the offset head with L1 Loss. We also add a detection head to the image features that predict the center of the 2D bounding boxes, helping to guide the features before we project them to voxel space.

For tracking, we predict the motion of each detection in the BEV space. Similar to the offset, we learn the offset to the location in the previous frame. As these offsets can vary in magnitude, we choose the Smooth L1 Loss.

4. Experiments

4.1. Datasets

Wildtrack [7] comprises real-world footage obtained from seven synchronized and calibrated cameras. These cameras capture an overlapping field-of-view of a 12×36 meter public area where pedestrian movement is unscripted. Ground plane annotations are offered on a 480×1440 grid, equating to 2.5 cm grid cells. On average, each frame contains 20 pedestrians covered by 3.74 cameras. The video is recorded at a resolution of 1080×1920 pixels at a frame rate of 2 fps.

MultiviewX [27] is a synthetic dataset modeled close to the specification of Wildtrack dataset using a game engine. This dataset includes views from 6 virtual cameras with an overlapping field-of-view encompassing a slightly smaller area (16×25 meters compared to 12×36 meters in Wildtrack). Annotation are provided on a ground plane grid of size 640×1000, with each grid representing the same 2.5 cm squares. With an average of 40 pedestrians per frame

	Lifting Method	Wildtrack				MultiviewX			
		MODA	MODP	Precision	Recall	MODA	MODP	Precision	Recall
DeepMCD [6]	Learned	67.8	64.2	85	82	70.0	73.0	85.7	83.3
Deep-Occlusion [1]	Learned	74.1	53.8	95	80	75.2	54.7	97.8	80.2
MVDet [27]	Persp. Proj.	88.2	75.7	94.7	93.6	83.9	79.6	96.8	86.7
SHOT [42]	Persp. Proj.	90.2	76.5	96.1	94.0	88.3	82.0	96.6	91.5
3DROM[†] [37]	Persp. Proj.	91.2	76.9	95.9	95.3	90.0	83.7	97.5	92.4
MVDeTr [26]	Persp. Proj.	91.5	**82.1**	97.4	94.0	93.7	91.3	**99.5**	94.2
EarlyBird [44]	Persp. Proj.	91.2	81.8	94.9	96.3	94.2	90.1	98.6	95.7
MVTT [31]	Persp.+RoI	**94.1**	81.3	**97.6**	**96.5**	95.0	**92.8**	99.4	95.6
TrackTacular	Persp. Proj.	91.8	79.8	96.2	95.6	95.9	89.2	**99.5**	96.4
	Bilin. Sampl.	92.1	76.2	97.0	95.1	**96.5**	75.0	99.4	**97.1**
	Depth Splat.	93.2	77.5	97.3	95.8	96.1	90.4	99.0	**97.1**
	Deform. Attn.	78.4	73.1	93.8	84.0	94.4	73.1	98.6	95.8

Table 1. **Pedestrian Detection.** Comparison of the BEV detection performance with the state-of-the-art methods on the Wildtrack and MultiviewX datasets. [†] 3DROM results are without additional data augmentations.

and coverage of 4.41 cameras per location, camera resolution and frame rate are the same as in the Wildtrack dataset. Like Wildtrack, the dataset has a length of 400 frames.

Synthehicle [24] is a synthetic dataset modeling intersections cameras for intelligent cities in CARLA [11]. 3-8 cameras record each intersection with a large overlapping area in the center of the intersection. The dataset models day, dawn, night, and rain conditions. The scenes are annotated per camera with a camera calibration. We consider the classes cars, trucks, and motorbikes. The dataset has separate towns for the test set, which allows for evaluating unseen intersections.

4.2. Evaluation Metrics

There are different philosophies on how to evaluate 3D detection. Three common protocols are used: 3D bounding boxes [17, 48], 2D BEV bounding boxes[17], 2D BEV center points [5]. We follow the 2D BEV center point protocol as it is the most common protocol for MTMC tracking, and it is more forgiving of minor errors in 3D detection.

Detection. Pedestrian detection is classified as true positive if it is within a distance $r = 0.5$ meter, which roughly corresponds to the radius of a human body. Following previous works [7, 27], 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. Aligning with our detection philosophy, we evaluate all tracking metrics on 2D BEV center points [5, 7]. We report the common MOT metrics [4, 48] and identity-aware metrics [39]. The threshold for a positive assignment is set to $r = 1$ meter. The primary metrics under consideration are Multiple Object Tracking Accuracy (MOTA)

	Wildtrack				
	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓
KSP-DO [7]	73.2	69.6	61.5	28.7	25.1
KSP-DO-ptrack [7]	78.4	72.2	60.3	42.1	14.6
GLMB-YOLOv3 [35]	74.3	69.7	73.2	79.5	21.6
GLMB-DO [35]	72.5	70.1	63.1	**93.6**	22.8
DMCT [53]	77.8	72.8	79.1	61.0	4.9
DMCT Stack [53]	81.9	74.6	78.9	65.9	4.9
ReST[†] [9]	86.7	84.9	84.1	87.8	4.9
EarlyBird [44]	92.3	89.5	**86.6**	78.0	4.9
MVFlow [12]	93.5	91.3	–	–	–
TrackTacular (Perspective Transform)	94.2	89.6	81.7	87.8	4.9
TrackTacular (Bilinear Sampling)	**95.3**	**91.8**	85.4	87.8	4.9
TrackTacular (Depth Splatting)	93.6	90.2	84.2	82.9	4.9
TrackTacular (Deformable Attention)	88.0	82.2	78.9	75.6	4.9
	MultiviewX				
	IDF1↑	MOTA↑	MOTP↑	MT↑	ML↓
EarlyBird [44]	82.4	88.4	86.2	82.9	**1.3**
TrackTacular (Perspective Transform)	84.2	91.4	**86.7**	85.5	2.6
TrackTacular (Bilinear Sampling)	**85.6**	**92.4**	80.1	**92.1**	2.6
TrackTacular (Depth Splatting)	83.4	91.8	84.7	90.8	2.6
TrackTacular (Deformable Attention)	84.8	91.4	80.6	89.5	2.6

Table 2. **Pedestrian Tracking.** Evaluation of tracking results on the Wildtrack and MultiviewX. [†]Re-computed by us.

		Scene Specific			Cross Scene		
Synthehicle	Lifting Method	IDF1↑	MOTA↑	MOTP↑	IDF1↑	MOTA↑	MOTP↑
TrackTacular	Bilinear Sampling	48.0	10.6	32.8	18.3	-22.0	29.3
	Depth Splatting	57.2	33.1	41.9	24.2	-1.5	32.4

Table 3. **Synthehicle.** Evaluation on the scene specific validation set and the cross scene test set. The validation set consists of temporally separated scenes from the train set and the test set contains only unseen scenes.

and IDF1. MOTA takes missed detections, false detections, and identity switches into account. IDF1 measures missed detections, false positives, and identity switches.

4.3. Implementation Details

Following [20, 26], 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 $\boldsymbol{t}$ of the camera extrinsic to avoid overfitting the decoder. We train the detector using an Adam optimizer with a one-cycle learning rate scheduler and a maximum learning rate of 10^{-3}. Depending on the size of the encoder, a batch size of $1-2$ is employed. To stabilize training, we accumulate gradients over multiple batches before updating the weights to have an adequate batch size of 16. The encoder and decoder network are initialized with weights pre-trained on *ImageNet-1K*. We run all experiments on RTX 3090 GPU.

Temporal Caching. Our method incorporates the BEV feature of the previous frame. A challenge is to have the previous feature cached while still achieving a high variation of batch samples, as the gradients are updated in order of the sequence. For testing, we can resort to a batch size of one to always have the previous feature computed, but for training, this would harm the performance due to a small batch size or slight sample variation. Thus, we build a custom sampler that composes batches according to the accumulated batch size in a semi-sequential order.

4.4. Results

Pedestrian. We report the detection results on both pedestrian datasets in Tab. 1. Compared to previous works, our approach can improve the state-of-the-art further. Only the two-stage approach MVTT [31] is better in MODA on Wildtrack. Our approach dominates all other metrics. The overall high values indicate that the results start to saturate. The results on MultiviewX improve with larger margins compared to Wildtrack. These results may indicate that the labeling accuracy on Wildtrack limits us as they annotated using perspective transform [7] and underlines the need for more challenging datasets.

All recent related work used the perspective projection as the lifting method, and our approach can outperform those approaches due to the additional temporal information. Overall, the parameter-free lifting methods, bilinear sampling, show competitive results for parameterized depth splatting. This observation aligns with the expectation that the depth splatting method can provide dense features throughout the observed space for this small area. The deformable attention method could perform better on our task, even though [20] showed it to be the most robust lifting method for autonomous driving tasks. Compared to the other methods, we observed strong overfitting effects during training, indicating that this method is not viable for small datasets like the one tested here.

	Detection		Tracking	
	MODA	MODP	IDF1	MOTA
Baseline (Bilin. Sampl.)	95.4	89.8	81.5	90.0
+ History Fusion	96.5	75.0	83.8	87.9
+ Motion Prediction	”	”	85.6	92.4

Table 4. **Temporal Ablation.** Evaluation of temporal aggregation components introduced by our approach compared to the baseline on the MultiviewX dataset.

We compare the tracking results in Tab. 2. Our approach improves the state-of-the-art (SOTA) on both datasets. Our work can further improve the tracking quality compared to the three most recent approaches. It shows that tracking in BEV is currently the most potent approach, as EarlyBird [44], MVFlow [12], ReST [9], and our approach follows this idea. EarlyBird [44] is most similar to our approach as it is also a one-shot tracker. However, it uses an appearance-based association and only the information of a single frame. The strength of our association method is mainly reflected in the IDF1 score, and the significant gain shows the advantages of our early-fusion tracker, which can combine appearance and motion cues. Overall, we can observe similar trends to those of the detection task as the improvements stagnate on Wildtrack.

Vehicle Results. In Tab. 3, we report the results on Synthehicle [24] with the two lifting methods. The main advantage of evaluation on Synthehicle is that we can evaluate scenes known during training (scene-specific) and new scenes (cross-scene). The much more complex dataset

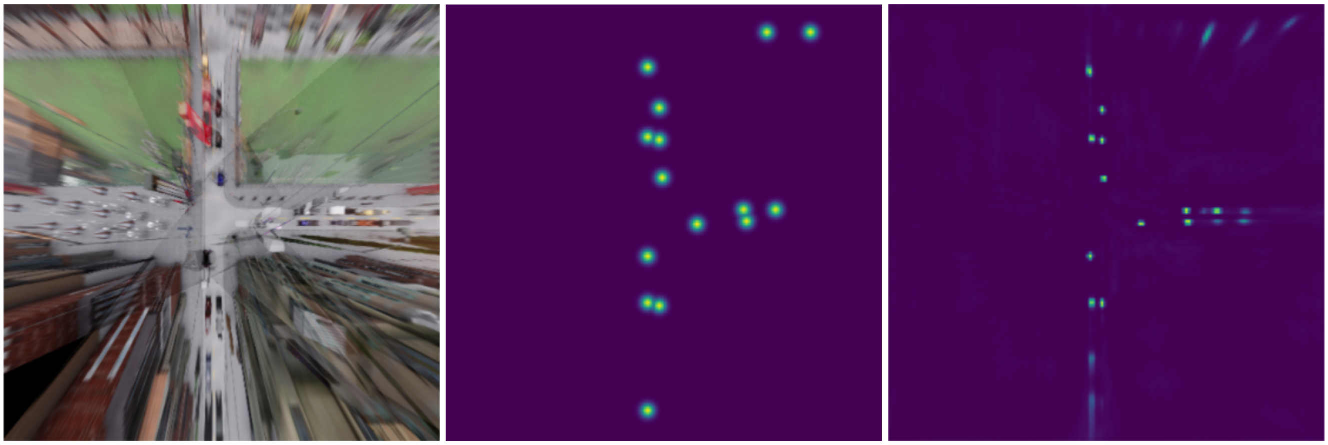

(a) Projected BEV (b) Ground Truth (c) Prediction

Figure 4. **Qualitative Results.** Detection example shown on Synthehicle. (a) shows the input images projected to the BEV space, (b) shows the ground truth heatmap of all vehicles, and (c) our prediction with bilinear sampling.

shows much lower tracking scores. The parameterized depth splatting lifting method outperforms the parameter-free bilinear sampling by a significant margin. Both approaches do not show robust scene generalization capabilities, mainly in the detection quality, as indicated by the low MOTA score.

Temporal Aggregation. The ability to model temporal information is crucial for a tracking model. In Tab. 4, we ablate the aggregation components of our model. The baseline is our proposed method without access to the history and motion prediction. With additional history frames, the detection accuracy increases, but the precision decreases. This obeservation indicates that the model can detect more pedestrians with history frames, but the location precision decreases due to ambiguity introduced with the history frames. The additional motion prediction only affects the tracking results, and the results improve significantly from our prediction.

4.5. Qualitative Results

In Fig. 4, we show an example from the Synthehicle validation set. First, in 4a, we projected all camera views perspectively to the ground plane to give an approximation of the BEV of the scene. Thus, it also approximates how the image features after the encoder is projected. The image features, or pixels, are stretched further on the outer parts of the scene. In 4b, we show all objects in the ground truth. Compared to the prediction (4c) of our approach with the bilinear lifting method, we show promising results in the center of the scene. The strength of the prediction declines on the outer parts of the scene. The predictions also show vehicles in areas not highlighted in the ground truth but visible in the projected BEV-View. Synthehicle builds the label based on 2D views and thus might miss detections in 3D.

5. Conclusion

Our paper gives an extensive overview of different lifting strategies for MTMC task. Combined with a motion-based tracking approach, we show SOTA results on two pedestrian datasets. Our results show that the results are saturating on Wildtrack and MultiviewX, requiring new datasets. Thus, we extended our evaluation to a roadside perception dataset. This dataset allowed for a new challenge in this area: scene generalization. However, all datasets considered still focus on 2D detections, and our results show that the field needs new 3D-first datasets as a standard benchmark. We are confident that our approach inspires a new MTMC dataset and new one-shot multi-view detection and tracking approaches.

References

[1] Pierre Baqué, François Fleuret, and Pascal Fua. Deep occlusion reasoning for multi-camera multi-target detection. In *ICCV*, pages 271–279, 2017. 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. 3

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *CVPR*, pages 941–951, 2019. 3, 4

[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] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 6

[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] 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. 1, 5, 6, 7

[8] 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

[9] 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. 1, 3, 6, 7

[10] Christian Creß, Walter Zimmer, Leah Strand, Maximilian Fortkord, Siyi Dai, Venkatnarayanan Lakshminarasimhan, and Alois Knoll. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970. IEEE, 2022. 1

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 6

[12] Martin Engilberge, Weizhe Liu, and Pascal Fua. Multi-view tracking using weakly supervised human motion prediction. In *WACV*, 2023. 6, 7

[13] 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

[14] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, pages 3038–3046, 2017. 3, 4

[15] 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. 1

[16] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE TPAMI*, 30(2):267–282, 2007. 3

[17] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 6

[18] Derek Gloudemans, Yanbing Wang, Gracie Gumm, William Barbour, and Daniel B Work. The interstate-24 3d dataset: a new benchmark for 3d multi-camera vehicle tracking. *arXiv preprint arXiv:2308.14833*, 2023. 1

[19] Adam W Harley, Shrinidhi K Lakshmikanth, Fangyu Li, Xian Zhou, Hsiao-Yu Fish Tung, and Katerina Fragkiadaki. Learning from unlabelled videos using contrastive predictive neural 3d mapping. *arXiv preprint arXiv:1906.03764*, 2019. 2, 5

[20] 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. 2, 3, 5, 7

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 4

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

[23] Tong He, Pei Sun, Zhaoqi Leng, Chenxi Liu, Dragomir Anguelov, and Mingxing Tan. Lef: Late-to-early temporal fusion for lidar 3d object detection. *arXiv preprint arXiv:2309.16870*, 2023. 5

[24] 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, 2023. 1, 6, 7

[25] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In *CVPR*, pages 3650–3657, 2013. 1, 3

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

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

[28] 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

[29] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960. 4

[30] 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. 3

[31] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Multi-view target transformation for pedestrian detection. In *WACV Worksh.*, pages 90–99, 2023. 2, 6, 7

[32] 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. 2, 3, 5

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, pages 2980–2988, 2017. 5

[34] 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. 1, 3

[35] 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

[36] 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. 2, 3, 4

[37] 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, 6

[38] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[39] 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

[40] 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

[41] 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. 3

[42] 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. 2, 6

[43] 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. 1

[44] Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. EarlyBird: Early-fusion for multi-view tracking in the bird's eye view. In *WACV Worksh.*, pages 102–111, 2024. 1, 6, 7

[45] 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

[46] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 2

[47] 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, 4

[48] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020. 6

[49] 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

[50] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing HyDRa: Hybrid fusion, depth consistency and radar for unified 3d perception. *arXiv preprint arXiv:2403.07746*, 2024. 2

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

[52] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, pages 11784–11793, 2021. 2

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

[54] 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. 3

[55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 4, 5

[56] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, pages 474–490. Springer, 2020. 3, 4