STCrowd: A Multimodal Dataset for Pedestrian Perception in Crowded Scenes

Peishan Cong[1], **Xinge Zhu**[2], **Feng Qiao**[3], **Yiming Ren**[1], **Xidong Peng**[1], **Yuenan Hou**[4], **Lan Xu**[1,7], **Ruigang Yang**[5], **Dinesh Manocha**[6], **Yuexin Ma**[1,7†]
[1]ShanghaiTech University [2]The Chinese University of Hong Kong
[3]RWTH Aachen University [4]Shanghai AI Laboratory
[5]University of Kentucky [6]University of Maryland at College Park
[7]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[1]{congpsh, renym1, pengxd, mayuexin}@shanghaitech.edu.cn

Abstract

Accurately detecting and tracking pedestrians in 3D space is challenging due to large variations in rotations, poses and scales. The situation becomes even worse for dense crowds with severe occlusions. However, existing benchmarks either only provide 2D annotations, or have limited 3D annotations with low-density pedestrian distribution, making it difficult to build a reliable pedestrian perception system especially in crowded scenes. To better evaluate pedestrian perception algorithms in crowded scenarios, we introduce a large-scale multimodal dataset, STCrowd. Specifically, in STCrowd, there are a total of 219 K pedestrian instances and 20 persons per frame on average, with various levels of occlusion. We provide synchronized LiDAR point clouds and camera images as well as their corresponding 3D labels and joint IDs. STCrowd can be used for various tasks, including LiDAR-only, image-only, and sensor-fusion based pedestrian detection and tracking. We provide baselines for most of the tasks. In addition, considering the property of sparse global distribution and density-varying local distribution of pedestrians, we further propose a novel method, Density-aware Hierarchical heatmap Aggregation (DHA), to enhance pedestrian perception in crowded scenes. Extensive experiments show that our new method achieves state-of-the-art performance for pedestrian detection on various datasets. https://github.com/4DVLab/STCrowd.git

1. Introduction

Accurate pedestrian perception in 3D space plays a crucial role in thorough scene understanding. Many applications also benefit from reliable and accurate pedestrian perception [2, 4, 61], including surveillance, serving robots, autonomous driving, *etc*. However, pedestrian perception is intractable for three reasons. First, pedestrians are not rigid bodies and they can have various poses. Second, humans are relatively small for sensors to capture compared with other agents, such as vehicles. For instance, in LiDAR point clouds, pedestrians in the distance are usually represented as a few sparse points. Third, people tend to congregate when walking, which makes the detection of each individual person harder. Occlusion in crowded scenarios is a thorny problem for pedestrian perception.

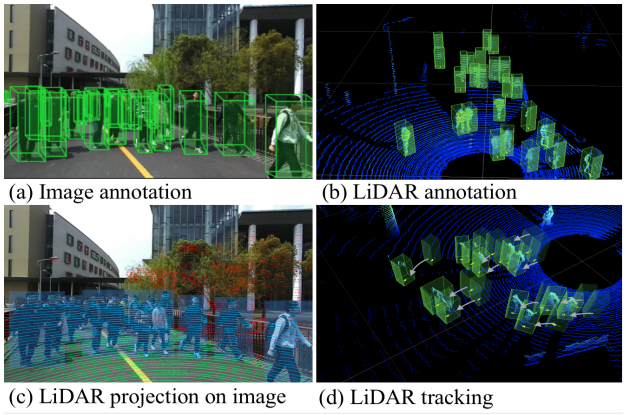


Figure 1. STCrowd provides 2D/3D image annotations, 3D point cloud annotations, and joint annotations for consecutive frames. Note that STCrowd contains a large quantity of crowded scenes with severe occlusions, which pose great challenges to pedestrian detection and tracking.

Many datasets [4, 9, 14, 16, 19, 23, 32, 36, 39, 53, 54] have been collected to accelerate the research on the pedestrian perception field. Previous pedestrian perception datasets can be classified into two groups: image-based pedestrian datasets and multimodal traffic datasets. The former [5–7, 9, 14, 16, 23, 32, 36, 52–54] focus on pedestrian detection and tracking on 2D images and merely provide

†: Corresponding author.

2D bounding box annotations, which is insufficient for deep models to infer accurate 3D positions of the pedestrians. The urgent demand for precise pedestrian perception in 3D space has given rise to a suit of 3D annotated datasets [4, 8, 19, 24, 33, 39, 58]. However, these datasets all focus on the traffic scenes, where most objects on the roads are vehicles and pedestrians are distributed sparsely, which limits the exploration and evaluation of learning-based perception methods, especially for crowded scenes.

Specifically targeting 3D pedestrian perception in challenging crowded scenarios, we introduce a large-scale multimodal dataset, STCrowd, with manually labeled 3D annotations for both images and point clouds. There are a total number of 219 K pedestrian instances in STCrowd with 20 persons per frame on average and more than 30 persons per frame in extremely crowded scenes. Specifically, there are 8 pedestrians in 5 meters on average centered on each person, which is much denser than contemporary 3D detection benchmarks, *e.g.*, nuScenes [4] and KITTI [19]. Due to the lack of crowded 3D pedestrian datasets, perception algorithms always suffer from severe occlusions when dealing with crowded scenarios. STCrowd is very useful for exploring more effective methods and testing their robustness. In addition, we capture the data in 9 different scenes, covering different weather, light conditions and road conditions. With rich annotations, STCrowd is applicable for different tasks, including LiDAR-only, image-only, and sensor-fusion based detection, tracking and even trajectory prediction. We also provide baselines for most of the tasks in this paper to facilitate further research.

For LiDAR-captured outdoor scenes, pedestrians typically account for a small portion of the whole scene. For crowded scenarios, pedestrians gather together, which causes different degrees of occlusion and makes it difficult to distinguish each individual pedestrian accurately in the crowd. Considering the sparse global distribution and density-varying local distribution of the pedestrians, we propose a novel method, Density-aware Hierarchical heatmap Aggregation (DHA), to enhance pedestrian perception especially in crowded scenes. Specifically, DHA is comprised of the spatial attention module and the hierarchical heatmap aggregation module. The former makes the network focus on the pertinent foreground regions and the latter helps distinguish individuals in density-varying scenes via multi-level heatmaps. We evaluate our method on STCrowd and achieve state-of-the-art performance. The extension of the proposed DHA to tracking problem and ablation studies on various backbones also demonstrate its effectiveness and good generalization capability. Our contribution is summarized as below:

1) We propose a large-scale multimodal pedestrian-oriented dataset in crowded scenarios with 3D manual annotations. High-density distributions of pedestrians result in severe occlusion, which bring challenges for accurate perception.
2) Our dataset can be used for various tasks, including LiDAR-only, image-only, and sensor-fusion based pedestrian detection and tracking. We provide baselines and metrics for most of the tasks to facilitate further research.
3) We propose a novel method to enhance LiDAR-based pedestrian perception in crowded scenes and achieve state-of-the-art performance on the STCrowd benchmark.

2. Related Work

Image-based Datasets Many datasets have been proposed over the last decade for pedestrian detection. Early datasets, like INRIA [15], ETH [18], TUD-Brussels [44], and Daimler [17] are too small for the training and generalization of deep learning-based methods. Caltech [16] is a widely-used pedestrian dataset with plenty of annotations. After that, more and more datasets were proposed for boosting data-driven human detection techniques, including KAIST [23], CityPersons [53], WiderPerson [54], NightOwls [32], etc. Especially, CrowdHuman [36] and DensePeds [5] provide many crowded scenes. In addition, there are also some datasets with traffic scenes containing pedestrians, such as CamVid [3], Vistas [31] Cityscapes [14], D^2-City [9], BDD100k [52], METEOR [7], etc. Note that all of them annotate pedestrians with 2D annotations, which is not applicable for real-world 3D perception.

Multimodal Datasets With the rapid development of autonomous driving, there is a growing demand for large-scale datasets with 3D annotations. Apolloscape Detection/Tracking [30] is a challenging urban traffic scenes dataset. It only has 3D annotations for the LiDAR point cloud. Because almost all the autonomous vehicles have both cameras and LiDAR sensors for perception, there are many multimodal datasets containing synchronized images and LiDAR point cloud, such as KITTI [19], nuScenes [4], Waymo Open [39], H3D [33], A2D2 [20], KAIST [13], A*3D [58], Argoverse [8], Lyft L5 [24], etc. However, they all focus on the driving scenes, where vehicles account for most of the scene, and pedestrians are distributed sparsely. In fact, pedestrians have free rotations and diverse poses, and they are much smaller than vehicles, which dramatically increases the difficulty in the detection and tracking. Furthermore, high-density crowd scenes are much more challenging due to the existence of severe occlusion. Our dataset provides diverse scenes with various densities and distributions of pedestrians, which is significant for testing perception methods' generalization.

3D Detection and Tracking LiDAR-only-based 3D detection methods aim to classify and locate the 3D bounding boxes in the given point cloud. Most of them [10, 34, 35, 38, 41, 47, 56, 57, 59, 60] first project the point cloud into a 3D or 2D representation, such as voxel and pillar. After that, the standard 2D convolution and 3D convolution are uti-

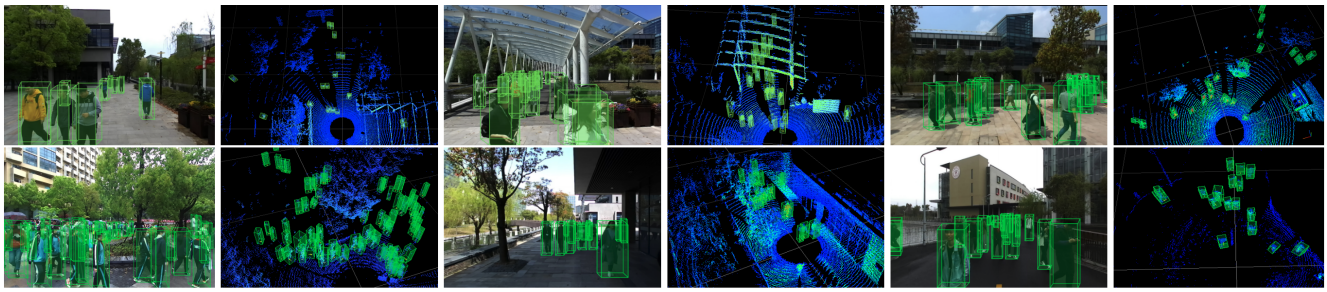
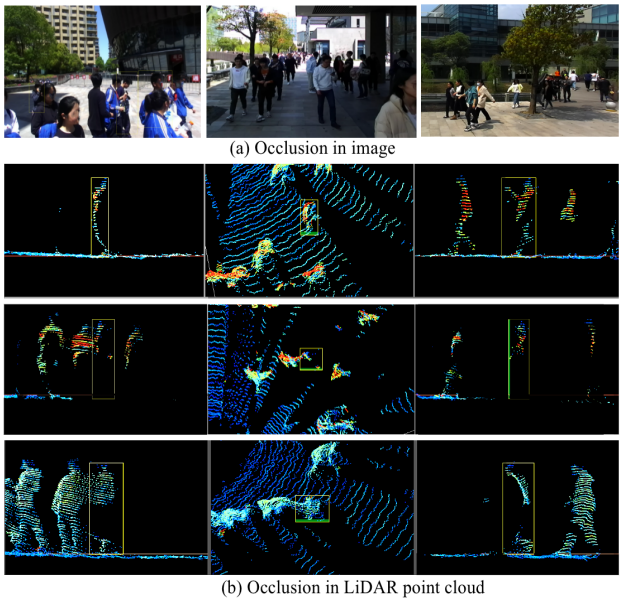


Figure 2. Scene examples of STCrowd with different backgrounds and weather, including clear weather (the first row), cloudy and rainy conditions (the second row). Note that the last figure shows the poor captured LiDAR point cloud due to the rainy conditions.

lized to process these structured representations. Another group of existing methods [37, 48, 49] aims to process the point cloud in the raw data, which better preserves the 3D geometric information but has a high computational cost for the large-scale point clouds. Because of the complementary roles of point clouds and images, the LiDAR and camera fusion methods have gained much attention in recent years. PointPainting [41] makes the sensor fusion in the point level with a hard-association. PointAugmenting [42] performs the point-wise fusion in the feature level. Furthermore, MV3D [11] and AVOD [26] perform fusion at the region proposal level. Similarly, [28, 29, 46, 51] project the point cloud onto the bird's eye view (BEV) and then fuse the image features in the BEV level. For the 3D tracking [22], existing methodologies for 2D tracking can be easily adapted to 3D space [2, 45]. A common pipeline is to combine the 3D detectors and 3D Kalman filters to perform 3D tracking [12, 43]. However, most of them often struggle with the challenging crowded scenes with severe occlusions. Using spatial attention and hierarchical heatmaps, our method can focus on pedestrians in large-scale scenes and distinguish individuals well for density-varying crowds.

3. Dataset

STCrowd is collected by a 128-beam LiDAR and a monocular camera, which are synchronized and mounted at a fixed position on the vehicle. The detailed set-up of the sensors is shown in the supplementary materials. The annotated dataset is comprised of 84 sequences and the total number of frames is 10, 891. Each sequence contains a variable number of continuously recorded frames, ranging from 50 to 800. There are 219 K and 158 K instance-level bounding box annotations in point clouds and images, respectively. Joint annotations of point clouds and images in sequences are also provided. In particular, we get official permission for collecting the data and we protect personal privacy by blurring faces shown in images. The data annotation project involved 20 people with professional skills and took 960 man-hours effort. And we have two rounds of quality inspection for each batch of data.



(a) Occlusion in image

(b) Occlusion in LiDAR point cloud

Figure 3. Occlusion cases in STCrowd. (a) shows occlusion cases in image. (b) demonstrates examples in LiDAR point cloud from different views (front, top, and side view) from left to right, in which severely occluded pedestrians are marked by yellow boxes.

3.1. Characteristics

Diverse scenes and weather. We collect data in different scenes and weather conditions as shown in Figure. 2. Our scenes include rich background with bridges, trees, buildings and designed architectures. Unlike traffic scenes in which pedestrians gather around junctions or roadsides, the distribution of pedestrians in our scenes is more diverse. The weather also varies to include clear, cloudy and rainy days. Different lighting conditions will influence the color information of the image, and the rainy or wet conditions will affect the reflection from the LiDAR sensor, resulting in fewer points on objects and the background (eg. for the last picture shown in Figure. 2, it is clear to see that the captured points are very limited), which is challenging for perception algorithms.

Diverse crowd densities. STCrowd contains crowd scenar-

Table 1. Comparison of STCrowd with popular multimodal datasets, where Fr denotes frames, PerFr denotes per frame of LiDAR point cloud, - represents unknown. Density-2/5/10 shows the average number of pedestrians within 2, 5, and 10 meters respectively centered on each pedestrian. Person/Range is the ratio of the number of pedestrians in each frame and the LiDAR scan diameter. The last three indicators measure the density of the dataset from different aspects.

	LiDAR Fr	3D Boxes	Beam	Person Num	Person PerFr	Person/Range	Density-2/5/10
PedX [25]	2.5k	0	-	14k	5.6	-	-
Argoverse [8]	22k	993k	32	110k	5	-	-
Lyft L5 [24]	46k	1.3M	64	210k	4.6	-	-
A*3D [58]	39k	230k	64	20k	0.5	-	-
KITTI [19]	15k	80k	64	4.5k	0.3	0.006	0.5/1.3/2.3
nuScenes [4]	40k	1.4M	32	208k	5	0.05	0.7/1.6/2.7
H3D [33]	27k	1M	64	280k	10	0.1	1.5/4.0/7.2
Waymo Open [39]	230k	12M	64	2.8M	12	0.16	1.0/2.9/5.6
ours	11k	219k	**128**	219k	**20**	**0.4**	**2.4/8.0/15.8**

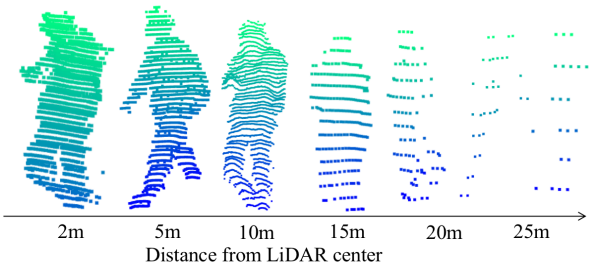


Figure 4. Wide-span point cloud densities for instances with the distance to LiDAR sensor increasing.

ios of various densities, which are divided into four levels, including fewer than 10, 10 $\sim$ 20, 20 $\sim$ 30, and more than 30 pedestrians. High-density is the most important characteristic of STCrowd. Table. 1 shows the comparison with related datasets, which are widely-used for the 3D perception of large-scale outdoor scenes. STCrowd is notable on three evaluation values. The first is the number of pedestrians per frame. Our dataset has 20 pedestrians on average, which obviously exceeds others. The second is the ratio of pedestrians and the scene range captured by LiDAR, which can show the density of the distribution of pedestrians in the whole scene. Statistics show that the value of our dataset is 2.5 times the density in Waymo and more than 4 times others. The last evaluation is to illustrate the degree of the crowd gathering by computing the average number of pedestrians in 2, 5, and 10 meters centered on each pedestrian. There are 2.4, 8, and 15.8 persons under such measurements for our dataset, which reveals the local high-density characteristic of STCrowd. When people gather, point clouds of different instances always stick with each other, which makes it difficult for perception methods to distinguish individuals accurately.

Dense crowds lead to severe occlusions in both images and LiDAR point cloud. As shown in Figure. 3, many pedestrians only have a partial body or only one arm or a head, which makes accurate perception difficult due to limited partial features. We annotate occlusion labels (from 0 to 2) for each pedestrian, measuring how much it is occluded, where 0, 1, and 2 denote none of the body, no more than half the body and over half the body occluded, respectively. Hierarchically dividing the dataset according to occlusion situations can help test the performance of methods in dealing with challenging cases.

Besides various scene-level densities, we also demonstrate diverse instance-level densities. As shown in Figure. 4, the point clouds of pedestrians become sparser as the distance to the LiDAR sensor increasing. Long-distance instances are hard to detect because the shape and scale information may loss on sparse points.

Diverse human poses. Our dataset has a diversity of human poses. Figure 5 shows some examples, like walking in person or in group, running, taking bicycles, taking balance cars, sitting, holding an umbrella, *etc*. The diversity in pedestrian poses further increases the difficulty of accurate perception.

3.2. Annotations

We provide high-quality manually labeled ground truth for both LiDAR point clouds and images. For annotations in point clouds, we labeled each pedestrian using a 3D bounding box $(x, y, z, l, w, h, \theta)$, where x, y, z denotes the center coordinates and l, w, h are the length, width, and height along the x-axis,y-axis and z-axis, respectively. Pedestrians with fewer than 15 points in the LiDAR point cloud are not annotated. For annotations of images, besides 3D bounding box, we also label the 2D bounding box with x, y, w, h for general 2D detection and tracking. For the objects captured by both the camera and LiDAR, we annotate the joint ID in sequences, which facilitates tracking and sensor-fusion tasks. The frequency of our annotation is 2.5HZ. We also provide annotations for the level of density and occlusion, which is mentioned above.

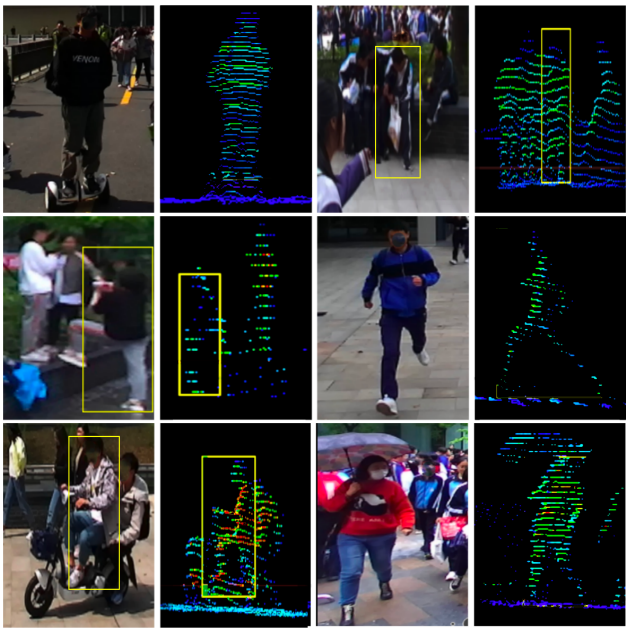


Figure 5. Diverse human poses in STCrowd. The same pedestrian in the image and the point cloud is marked by yellow boxes.

3.3. Tasks & Metrics

Our multi-modal dataset supports detection, tracking, and prediction tasks. We give evaluation metrics in this section to provide benchmarks on our dataset.

3.3.1 Detection Metric

Average Precision metric. Following [4], we use Average Precision (AP) metric with the 3D center distance threshold. For crowded scenes, the distance thresholds are chosen from $D = \{0.25, 0.5, 1\}$ meters and the mean Average Precision (mAP) is calculated by:

$$mAP = \frac{1}{|D|} \sum_{d \in D} AP_d$$

Average Recall with different occlusion levels. In addition to AP, for crowded scenes, the performance on occluded instances are also considered, and we calculate the average recall with different center distance thresholds $D = \{0.25, 0.5, 1\}$ for different levels of occlusion i:

$$AR_i = \frac{1}{|D|} \sum_{d \in D} Recall_{i,d}, i \in \{0, 1, 2\}$$

3.3.2 Tracking Metric

MOTA We use traditional Multi-Object Tracking Accuracy (MOTA) to measure the tracking result:

$$MOTA = 1 - (FP + IDS + FN)/GT$$

where FP and FN are false positive and false negative, IDS denotes the false ID matching for tracking in different timesteps, and GT is the number of ground truth tracked instances.

ML & MT ML(mostly loss) is the proportion of successful track matching of ground truth in less than 20% of the time in all tracking targets. MT(mostly track) is the proportion of successful track matching of ground truth in more than 80% of the time in all tracking targets.

3.3.3 Prediction Metric

FDE & MDE Final displacement error (FDE) is the Euclidean distance between the predicted output and the ground truth at the last time step and Mean displacement error (MDE) is the average Euclidean distance between the predicted output and the ground truth for each time step.

4. DHA: Density-aware Hierarchical Heatmap Aggregation

For pedestrian detection in crowded scenarios, pedestrians are always walking together, resulting in local high density and occlusion in both the point cloud and the image. Moreover, the density distribution and pose types of pedestrians vary. These challenges narrow down the capability of existing methods for accurate pedestrian detection in such conditions. To tackle these problems, we propose the density-aware hierarchical heatmap aggregation (DHA) module shown in Figure. 6, which makes the model learn the attention on the location of individuals and produces multi-scale predictions covering regions with different densities, in which the proposed module could mitigate too much background influence and tackle the problem of pedestrian clustering at various densities levels.

In what follows, we give detailed explanations for these two main components, *i.e.*, the spatial attention module focusing on the pertinent regions of pedestrians and the multi-level heatmap loss covering varying density conditions.

4.1. Spatial Attention Module

Crowds of pedestrians tend to be clustered, presenting locally high density and globally sparse distribution. Hence, attention to these foreground regions is crucial to the performance of pedestrian detection. To this end, we follow [40] to model the global attention with a transformer. As shown in Figure. 6, we apply the triplet <Query, Key, Value> attention layer to extract the correlation among different locations and reweight these locations. For the final output $\bar{X}$,

$$\bar{X} = \text{softmax}(QK^T) \times V,$$

where Q, K, V denote the output of Query, Key and Value layer.

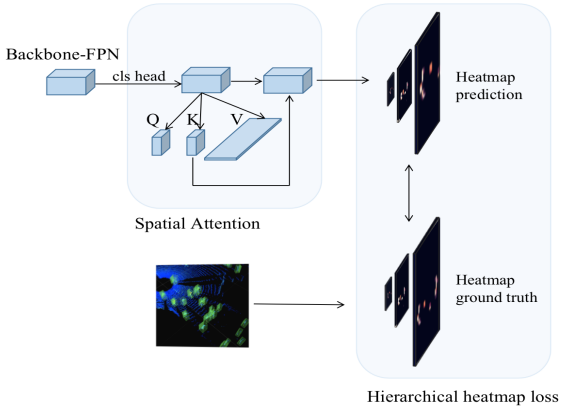


Figure 6. Density-aware hierarchical heatmap aggregation. We design the spatial attention with multi-level Gaussian score map supervision to tackle the problem of various density distribution and background influences, where it can also act as a plug-in for different backbones.

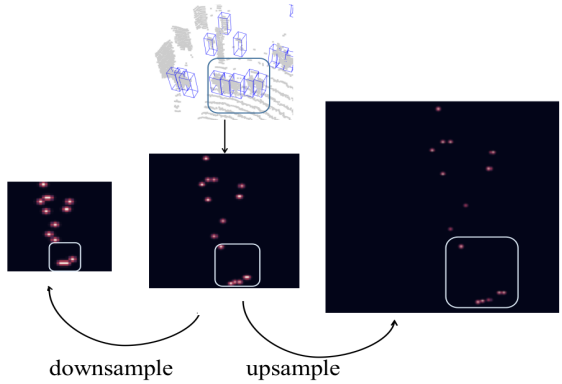


Figure 7. Hierarchical heatmaps. We design the spatial Hierarchical Gaussian score heatmap supervision to tackle the problem of various density distributions and background influences. The coarse-level heatmap is downsampled from the regular one, in which the positive regions occupy a larger portion targeting a balanced foreground / background ratio. For the fine-grained level heatmap, it can clearly distinguish the closer pedestrians for accurate one-to-one assignment (as shown in the rectangle).

4.2. Hierarchical Heatmaps

Balanced positive and negative samples also have a great impact on high-performance 3D detection. However, unlike big objects (like truck covering a large portion of heatmap), pedestrians only occupy a limited space, resulting in most of the heatmap being zero (or negative region). The density-varying property also worsens the condition. We thus introduce hierarchical heatmaps to construct the multi-level detection targets, where the coarse-level heatmap could balance the ratio of positive and negative samples and the fine-grained heatmap can better tackle the clustering pedestrians with an accurate one-to-one assignment (avoid one grid in the heatmap representing more than two persons).

Specifically, we use a modified CenterHead [50] to classify and localize the pedestrians. The ground truth heatmap is a Gaussian map produced based on 3D centers of annotated bounding boxes. As shown in Figure 7, the features generated from Spatial attention module are first upsampled to get the fine-grained feature maps and corresponding heatmaps, and then down-sampled to obtain the coarse-level features, where fine-grained level heatmap delivers an accurate one-to-one assignment for close and crowded pedestrians and coarse-level heatmap balances the ratio of foreground and background samples. Gaussian focal loss is calculated on each pair of prediction and target heatmaps.

With the cooperation of these two components, *i.e.*, spatial attention module and hierarchical heatmaps, DHA handles the crowd scenarios with varying density well.

5. Experiments

In this section, we first provide the detailed experimental setup, then evaluate extensive methods and our proposed DHA for 3D detection on STCrowd. Furthermore, we demonstrate the performance of DHA on 3D tracking and prove its generalization capability by ablation study. Finally, the benchmark of trajectory prediction on our dataset is provided to facilitate the research of crowd prediction. Moreover, we provide more analyses and qualitative results in the supplementary material.

5.1. Baseline

We present several popular baselines with different modalities for 3D detection. For image-based method, CenterNet [55] regresses 3D bounding boxes in the world-coordinate system from only monocular images. For LiDAR-only 3D detectors, anchor based and anchor-free methods are used to evaluate the performance on our dataset, including PointPillar [27] and CenterPoint [50]. Furthermore, we also evaluate two point encoding methods for CenterPoint, *i.e.*, Voxel-CenterPoint and Pillar-CenterPoint. Our method takes the Voxel-CenterPoint as the backbone and employs DHA as the classification head. For LiDAR-image-fusion-based 3D detectors, we show current SOTA PointPainting [41] and PointAugmenting [42], where pixel-wise prediction and pixel-wise image features are used to represent the image. Note that all the backbone for these fusion-based methods is the Voxel-CenterPoint.

5.2. Implementation Details

For experiments on STCrowd, we set the detection range to [0, 30.72m] for the X axis, [-20.48m, 20.48m] for Y axis, and [-4m, 1m] for Z axis. Voxel-CenterPoint employs a (0.12m, 0.16m, 0.2m) voxel size, and Pillar-CenterPoint and PointPillars utilize a (0.12m, 0.16m) grid. For the anchor-based method, the anchor is set as [0.57m, 0.6m,

Table 2. Benchmarks for image-only, LiDAR-only, and LiDAR-image-fusion-based 3D detection on validation set of STCrowd. AP(d) denotes that different meters are used as matching thresholds of 3D center distance d. AR_i represents the average recall on easy, moderate, and hard cases respectively with different occlusion levels i.

Methods	Modality	AP(0.25)	AP(0.5)	AP(1.0)	mAP	AR_0	AR_1	AR_2
CenterNet [55]+ResNet18	RGB	0.009	0.091	0.397	0.166	0.456	0.350	0.285
CenterNet [55]+ResNet101	RGB	0.056	0.112	0.486	0.203	0.478	0.361	0.273
CenterNet [55]+DLA34	RGB	0.041	0.200	0.467	0.236	0.578	0.451	0.349
PointAugmenting [42]	RGB+LiDAR	0.483	0.629	0.649	0.587	0.932	0.866	0.800
PointPainting [41]	RGB+LiDAR	0.509	0.638	0.656	0.601	0.929	0.867	0.783
PointPillar [27]	LiDAR	0.091	0.276	0.368	0.245	0.576	0.399	0.238
Pillar-Center [50]	LiDAR	0.456	0.574	0.592	0.541	0.866	0.811	0.706
Voxel-Center [50]	LiDAR	**0.505**	0.613	0.628	0.582	0.859	0.834	0.740
Ours	LiDAR	0.498	**0.667**	**0.685**	**0.617**	**0.902**	**0.873**	**0.782**

Table 3. Mean Average Precision (mAP) for 3D pedestrian detection on Waymo (W^{level}_{range}) [39], H3D [33], and nuScenes [4] datasets.

Dataset	W^1_{30}	W^2_{30}	W^1_{50}	W^2_{50}	H3D	nuScenes
Pillar-Center	0.697	0.652	0.535	0.472	0.478	0.719
Voxel-Center	0.701	0.649	0.598	0.532	0.595	0.783
Ours	**0.726**	**0.673**	**0.638**	**0.569**	**0.609**	**0.795**

1.7m] which is calculated as the average size of pedestrian 3D ground truth bounding boxes. For post-processing of detection results, we use a circle NMS method which keeps only one instance prediction within radiance fewer than 0.3m to reduce redundant bounding boxes and drops the predicted box which has fewer than 5 points.

5.3. Results

LiDAR-only 3D detection We compare the results of our proposed method with existing anchor-based and anchor-free methods in Table. 2. Specifically, the anchor-free methods perform much better than anchor-based methods. The various human poses and densities become a hindrance for anchor-based methods, making it difficult for these anchors to cover pedestrian locations well. Moreover, the proposed DHA achieves state-of-art performance compared with anchor-free methods, demonstrating that DHA can tackle the issue of density-varying and unbalanced samples well, while vanilla anchor-free methods do not well attend.

For crowded scenes, we show the result of average recall AR_i (as shown in Section 3.3.1) for different occlusion levels in Table.2 to evaluate the performance when facing key challenges in pedestrian detection. Consistently, the proposed method achieves about 4% and 6% improvement for each level compared with the Voxel-CenterPoint and Pillar-CenterPoint backbone, respectively.

We further provide the visualization results for the detection task in crowded scenes in Figure. 8. It can be observed that in crowded scenes, our method performs much better, even covering these challenging pedestrians when they are extremely close to each other and occluded severely (as shown in rectangles), while the baseline cannot distinguish these boxes clearly and misses some dense predictions.

We also conduct experiments in Table. 3 for 3D pedestrian detection on three other large-scale datasets. Obviously, our method consistently outperforms baselines (current SOTAs) on contemporary benchmarks, demonstrating good generalization capability of our method on 3D pedestrian detection.

Multimodal fusion-based 3D detection As shown in Table. 2, both fusion methods perform better than LiDAR-only baselines, which demonstrates the image features play a complementary role compared with LiDAR point clouds. Although our method only unitize LiDAR features, it is still comparable to sensor-fusion-based methods.

Image-only 3D detection It is obvious that there is a big gap between the image-only method and others. Because it is difficult to estimate the depth information from monocular images and severe occlusions of the crowd in images make things worse. In this manner, LiDAR point clouds provide an appealing option and act as a complementary function to tackle these occlusions to some extend.

5.4. Ablation studies

First, we perform ablation experiments to investigate the generalization ability of the proposed modules on various backbones. The results on the validation set are reported in Table 4 and we test our DHA module on pillar-based and voxel-based LiDAR-only-based detection backbones. The results demonstrate that it consistently improves the performance with a large margin, *i.e.*, 5% and 3.5%, re-

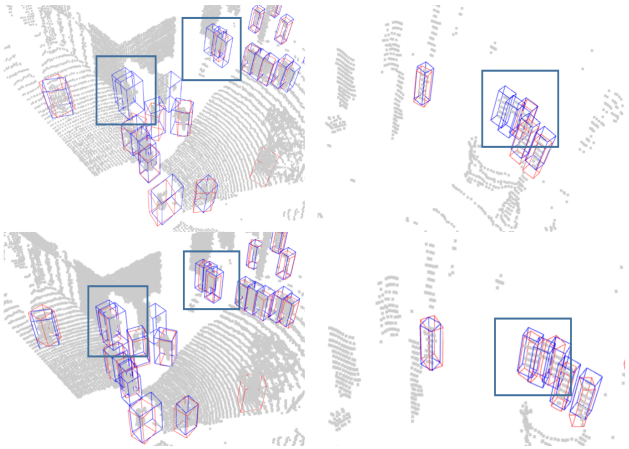


Figure 8. The detection visualization on crowded scenes. The first row is the prediction results from the baseline method [50] and the bottom shows our results. The blue boxes are ground truth, and the red boxes are predictions. It can be found that for some crowded regions and pedestrians (as shown in rectangles), the baseline method often omits and mismatches, while our method achieves better results on such cases because the proposed DHA can focus more on the foreground and distinguish the crowded regions with fine-grained heatmaps.

Table 4. Ablation studies for DHA block on different backbones on STCrowd validation set.

Pillar-Center	Voxel-Center	PointAugmenting	DHA	mAP
✓				0.541
✓			✓	0.591
	✓			0.582
	✓		✓	0.617
		✓		0.587
		✓	✓	0.594

Table 5. Ablation study of DHA on STCrowd.

Methods	ours	w/o SAM	w/o HH	Voxel-Center
mAP	0.617	0.601	0.603	0.582

spectively, compared to the baseline. We also test DHA on LiDAR-image-fusion-based detection backbone, PointAugmenting, and still get improvement (because image features already provide a remedy for the crowded scene, DHA only achieves a slight gain). We also conduct the ablation study for Spatial Attention Module (SAM) and Hierarchical Heatmaps (HH) (Table. 5). Results show the effectiveness of two modules in improving the performance.

5.5. 3D Tracking

For tracking tasks, we learn to predict a two-dimensional velocity estimation for each detected object as an additional regression output following the methodology of CenterPoint [50]. We mainly conduct experiments on CenterPoint [50] with pillar and voxel representation, respectively. The proposed DHA module is also incorporated to investigate its generalization ability on the tracking task (Table.6). It can be found that our proposed DHA consistently achieves the better results.

Table 6. Results of LiDAR point cloud tracking on validation set of STCrowd.

Methods	MOTA ↑	MT ↑	ML ↓
Pillar-Center	0.245	0.295	0.102
Voxel-Center	0.342	0.355	**0.084**
Ours	**0.368**	**0.363**	0.086

5.6. Trajectory Prediction

As shown in Table. 7, we also provide the baselines of trajectory prediction on our crowd dataset, including popular vanilla-LSTM [21], social-LSTM [1], and StarNet [62]. Our dataset can boost the research of action and trajectory prediction in crowd scenes by involving more multimodal inputs or features.

Table 7. Results on trajectory prediction task.

Methods	FDE ↓	MDE ↓
LSTM [21]	1.133	0.648
Social-LSTM [1]	1.122	0.638
StarNet [62]	0.983	0.404

5.7. More Applications

Accurate pedestrian perception leads to wider applications, like the on-campus delivery robots and intelligent patrols for stations, for which dense and crowded campus scene would be the major challenges. Our dataset can provide a benchmark and challenging metrics for them.

6. Conclusion

Focusing on 3D perceptions in crowded scenarios, we propose a new multimodal dataset with diverse crowd densities, multiple scenes, various weather, and different human poses. In particular, our dataset contain situations with high density and severe occlusions, which is challenging for current 3D perception methods. Based on multimodal data and annotations, our dataset can facilitate many perception tasks. Benchmarks on most of the tasks are provided in the paper. In addition, we propose a novel method to achieve more accurate perception on crowded scenes by considering the properties of pedestrian distribution. Experiments illustrate the superiority and generalization capability of our method.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016. 8

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019. 1, 3

[3] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008. 2

[4] 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. *CVPR*, 2020. 1, 2, 4, 5, 7

[5] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Densepeds: Pedestrian tracking in dense crowds using front-rvo and sparse features. In *IROS*, pages 468–475. IEEE, 2019. 1, 2

[6] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *CVPR*, pages 8483–8492, 2019. 1

[7] Rohan Chandra, Mridul Mahajan, Rahul Kala, Rishitha Palugulla, Chandrababu Naidu, Alok Jain, and Dinesh Manocha. Meteor: A massive dense & heterogeneous behavior dataset for autonomous driving. *arXiv preprint arXiv:2109.07648*, 2021. 1, 2

[8] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Sławomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. *CVPR*, pages 8740–8749, 2019. 2, 4

[9] Zhengping Che, Max Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D2-city: A large-scale dashcam video dataset of diverse traffic scenarios. *ArXiv*, abs/1904.01975, 2019. 1, 2

[10] Qi Chen, Lin Sun, Zhixin Wang, K. Jia, and A. Yuille. Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots. *ECCV*, 2020. 2

[11] Xiaozhi Chen, Huimin Ma, Jixiang Wan, B. Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. *CVPR*, 2017. 3

[12] Hsu-kuang Chiu, Antonio Prioletti, Jie Li, and Jeannette Bohg. Probabilistic 3d multi-object tracking for autonomous driving. *arXiv preprint arXiv:2001.05673*, 2020. 3

[13] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 2

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CVPR*, pages 3213–3223, 2016. 1, 2

[15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. Ieee, 2005. 2

[16] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2011. 1, 2

[17] Markus Enzweiler and Dariu M Gavrila. Monocular pedestrian detection: Survey and experiments. *TPAMI*, 31(12):2179–2195, 2008. 2

[18] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8. IEEE, 2008. 2

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32:1231 – 1237, 2013. 1, 2, 4

[20] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020. 2

[21] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 8

[22] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *CVPR*, pages 5390–5399, 2019. 3

[23] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, pages 1037–1045, 2015. 1, 2

[24] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Level 5 perception dataset 2020. 2019. 2, 4

[25] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2):1940–1947, 2019. 4

[26] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L. Waslander. Joint 3d proposal generation and object detection from view aggregation. *IROS*, 2018. 3

[27] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *CVPR*, 2019. 6, 7

[28] Ming Liang, Binh Yang, Yun Chen, Rui Hu, and R. Urtasun. Multi-task multi-sensor fusion for 3d object detection. *CVPR*, 2019. 3

[29] Ming Liang, Binh Yang, Shenlong Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. *ECCV*, 2018. 3

[30] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, 2019. 2

[31] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 4990–4999, 2017. 2

[32] Lukás Neumann, Michelle Karg, Shanshan Zhang, Christian Scharfenberger, Eric Piegert, Sarah Mistr, Olga Prokofyeva, Robert Thiel, Andrea Vedaldi, Andrew Zisserman, and Bernt Schiele. Nightowls: A pedestrians at night dataset. In *ACCV*, 2018. 1, 2

[33] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557, 2019. 2, 4, 7

[34] C. Qi, Xinlei Chen, O. Litany, and L. Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. *CVPR*, 2020. 2

[35] C. Qi, O. Litany, Kaiming He, and L. Guibas. Deep hough voting for 3d object detection in point clouds. *ICCV*, 2019. 2

[36] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1, 2

[37] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. *CVPR*, 2019. 3

[38] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *TPAMI*, 2021. 2

[39] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, and etc. Scalability in perception for autonomous driving: Waymo open dataset. *CVPR*, pages 2443–2451, 2020. 1, 2, 4, 7

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[41] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. *CVPR*, 2020. 2, 3, 6, 7

[42] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. *CVPR*, 2021. 3, 6, 7

[43] Xinshuo Weng and Kris Kitani. A baseline for 3d multi-object tracking. *arXiv preprint arXiv:1907.03961*, 1(2):6, 2019. 3

[44] Christian Wojek, Stefan Walk, and Bernt Schiele. Multi-cue onboard pedestrian detection. In *CVPR*, pages 794–801. IEEE, 2009. 2

[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

[46] Liang Xie, Chao Xiang, Zhengxu Yu, Guodong Xu, Zheng Yang, Deng Cai, and Xiaofei He. Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. *AAAI*, 2020. 3

[47] Binh Yang, Wenjie Luo, and R. Urtasun. Pixor: Real-time 3d object detection from point clouds. *CVPR*, 2018. 2

[48] Zetong Yang, Y. Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. *CVPR*, 2020. 3

[49] Zetong Yang, Y. Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. *ICCV*, 2019. 3

[50] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *CVPR*, 2021. 6, 7, 8

[51] Jin Hyeok Yoo, Yeocheol Kim, Ji Song Kim, and J. Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. *ECCV*, 2020. 3

[52] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *CVPR*, pages 2633–2642, 2020. 1, 2

[53] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, pages 3213–3221, 2017. 1, 2

[54] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z. Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia (TMM)*, 2019. 1, 2

[55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *ArXiv*, abs/1904.07850, 2019. 6, 7

[56] Yin Zhou, Pei Sun, Y. Zhang, Dragomir Anguelov, J. Gao, Tom Y. Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. *CoRL*, 2019. 2

[57] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *ArXiv*, 2019. 2

[58] Sijie Zhu, Taojiannan Yang, Mat'ias Mendieta, and Chen Chen. A3d: Adaptive 3d networks for video action recognition. *ArXiv*, abs/2011.12384, 2020. 2, 4

[59] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. *ECCV*, 2020. 2

[60] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, and etc. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, pages 9939–9948, 2021. 2

[61] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *TPAMI*, 2021. 1

[62] Yanliang Zhu, Deheng Qian, Dongchun Ren, and Huaxia Xia. Starnet: Pedestrian trajectory prediction using deep neural network in star topology. In *IROS*, pages 8075–8080. IEEE, 2019. 8