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

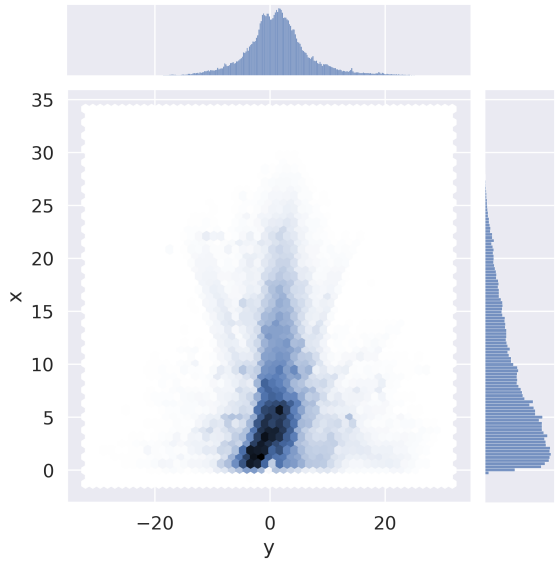

Figure 1. The pedestrian position distribution in BEV for point cloud. In STCrowd, the pedestrians distribute from 0 to 35 meters in x-axis and -30 to 30 meters in y-axis.

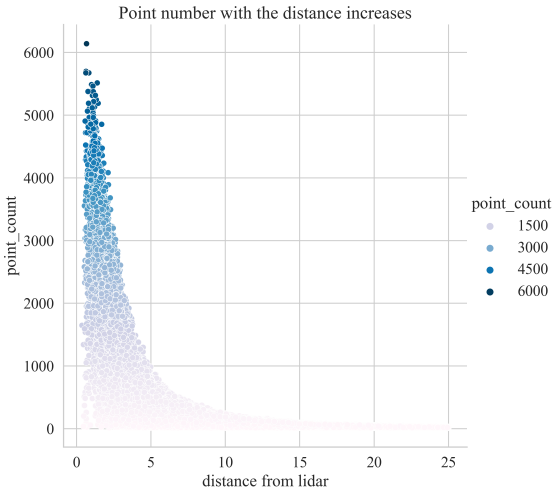

Figure 2. The number of points per instance decreases with the distance from LiDAR increases. The gradation of color represents the number of points.

1. Additional details of STCrowd

Sensor setup. The vehicle is equipped with following sensors: Zed Sterelab2 stereo camera(15Hz frame rate, 1280 × 720 resolution and 60 field-of-view(FOV) for left and right, 70 FOV for center) and OS0 3D LiDAR with 128 beams (More details are provided in Table 1).

Sensor synchronization. To achieve good cross-modal data alignment between different sensors, the timestamp of the LiDAR is the time when the full rotation of current frame is achieved and the correspond of timestamps for different devices is achieved by special posture when recording data. We keep the common frequency as 5 Hz and annotate the frames per 0.4 second.

Table 1. LiDAR specification.

Azimuthal FOV	360
Vertical FOV	90° (±45°)
Vertical resolution	0.7
Frequency	10Hz
range	55m
point per second	2,621,440

Pedestrian position distribution in BEV. We label the point cloud within the range of 180°. With the straight front of the camera as the reference direction, only the left and right 90° are labeled. The pedestrian position distribution in BEV for point cloud is presented in Fig 1.

Diverse instance-level densities and human poses . The density and position distribution is full of diversity which

contains both sparse and dense instances. We demonstrate diverse instance-level densities distribution in Figure 2 showing that the instance get sparser with fewer points as the distance from the LiDAR center increases.

Our dataset has a diversity of human poses. In addition to the cases shown in Figure 6 of the main content, we show more cases in Figure 3, such as walking alone or in group, walking with phone calls, taking skateboards, bowing, *etc*. The diversity in pedestrian poses further increases the difficulty of accurate perception.

Figure 3. Diverse human poses in STCrowd.

2. More qualitative results

2.1. Point cloud detection

As can be seen from Figure 4, our method makes fewer false predictions on the background than the backbone model owing to the spatial attention and hierarchical heatmap aggregation module (highlighted with black rounded rectangles).

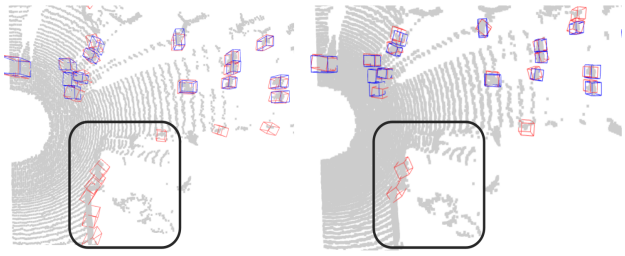

Figure 4. Visual comparison of the prediction of the backbone model (left) and our method (right).

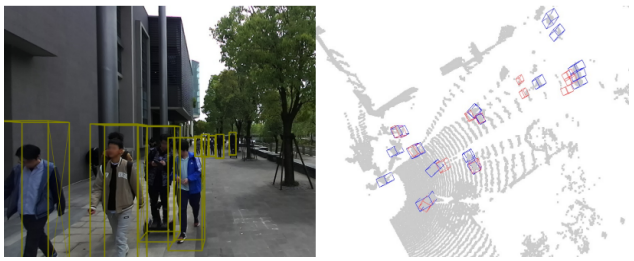

Figure 5. Visualization of the model prediction on the image-only detection. The left is the predicted 3D bounding boxes (yellow) on the image and the right is the result in point cloud (prediction is red and ground truth is blue). We can see that due to the challenge from monocular images depth prediction and detection, performance from images is poor when pedestrians are far from the LiDAR center.

Figure 6. Visual comparison of different backbones on the image-only detection. The predicted bounding boxes are in yellow. The first row is the prediction of ResNet-101 and the second row is the prediction of DLA-34. Due to its poor performance in depth prediction, the 3D projection results on the images can be misaligned, and DLA-34 performs slightly better than ResNet-101.

2.2. Image-only detection

We provide the image-only detection result in Figure 5 and Figure 6. Unlike binocular cameras, monocular cameras cannot obtain precise depth information from a single image. Accurate depth predictions are difficult with image-only information and severe occlusions of the crowd in images further bring challenges to monocular-image-based detection.