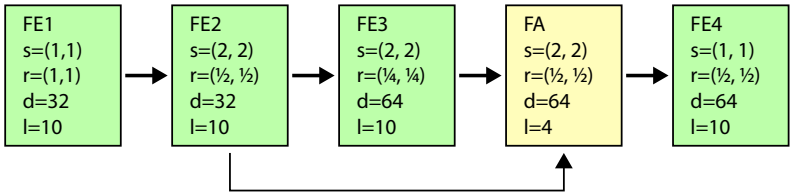
A. Additional Details on the Backbone

We use the basic building blocks proposed in [14]: the **feature extractor (FE)** and the **feature aggregator (FA)**. Figure 4 in [14] shows a detailed diagram. In words, the FE block consists of ten 3×3 filters. Every two layers are grouped and bypassed by a skip connection. Whenever a straightforward skip connection is not possible due to mismatched spatial resolution or depth, a 1×1 filter with potential striding is applied. The FA block is used to up-sample lower resolution feature maps back to a high resolution for skip connections. It first applies a transposed convolution filter to up-sample the lower resolution feature, which is concatenated to a skip connection from a high-resolution feature from a previous layer before down-sampling. The combined feature then undergoes 4 additional 3×3 filters.

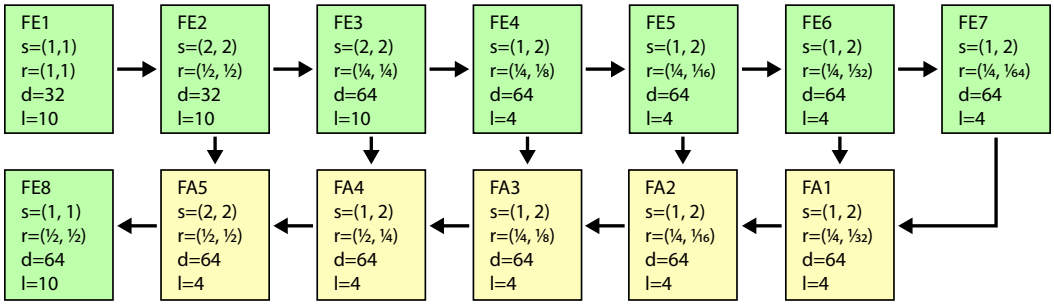
Fig. 4 shows backbone architectures for pedestrians and vehicles. Vehicles appear wider in the range image and require a larger receptive field. Therefore, the vehicle model is a lot deeper. The new feature extractors only have 4 instead of 10 convolutional layers each.

B. Additional Details on the Camera Backbone and Fusion

We use a U-Net [20] that resembles the 2D backbone network as depicted in Figure 4 of [33]. We make a small modification in that we skip the initial down-sampling by a stride of 2. The network has 16 convolutional layers with a kernel of 3×3 each, split into 3 blocks at an increasingly smaller resolution. Features from each of the 3 blocks are concatenated to create the final feature map.



(a) Backbone for pedestrians.



(b) Backbone for vehicles.

Figure 4: Backbone architecture for 3D pedestrian and vehicle detection. FE and FA are feature extractors and feature aggregators outlined in the text. **s** is stride in height and width and is applied at the start of each block. **r** is the relative resolution to the original input size after applying the stride. **d** is the number of channels for all convolutional layers inside the block. **l** is the number of 3×3 convolutional layers.

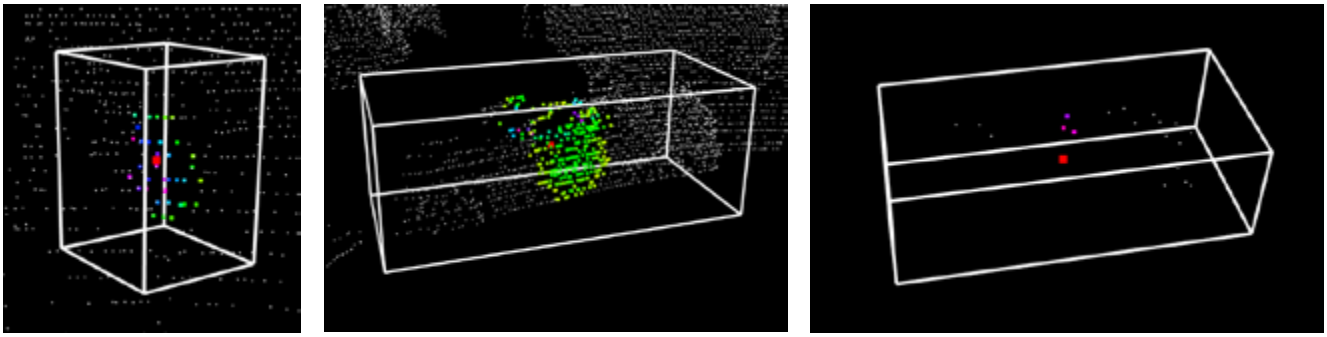


Figure 5: 3D CenterNet classification targets. White box denotes the groundtruth bounding box. Red dot is the center of the box. White points have the target value 0.0. Rainbow colors indicate the target values ranging from 0.1 (green) to 1.0 (purple). Left: pedestrian. Center: a close-by car with dense points. Right: a far away car with sparse points. Note that the closest points have the target value 1.0 despite being relatively far away from the center due to normalization.