## A. Additional details on SPFE

SPFE is composed of blocks illustrated in Fig. 4. PedL and CarL have been illustrated in Fig. 4. Architecture details of PedS, CarS and CarXL can be found in Fig. 8. PedS, PedL, CarS, CarL use 2D sparse convolutions and have channel size for all convolutions set to 96. CarXL use 3D sparse convolutions and have channel size for all convolutions set to 64. CarXL does not have PointNet within each 3D voxel.

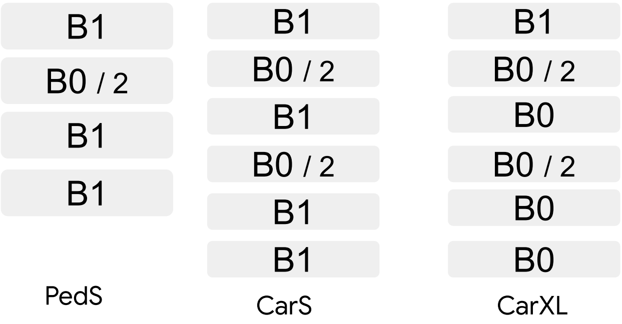

Figure 8. SPFE net architectures for CarS, PedS and CarXL.

## B. More Details on Temporal Fusion

1) Temporal RSN duplicates the RIFE (§3.1) and Foreground Point Selection part (§3.3) for each temporal frame. Shown in Fig. 9, each branch shares weights and matches the architecture for single frame RSN. These branches are trained together while during inference only the last frame is computed as other time-steps reuse previous results. 2) After segmentation branches, points are gathered to multiple set of points $P_{\delta_i}$ where $\delta_i$ is the frame time difference between frame $0$ (latest frame) and frame $i$ which is usually close to $0.1 * i$ seconds. Each point $\boldsymbol{p}$ in $P_{\delta_i}$ is augmented with $\boldsymbol{p} - \boldsymbol{m}$, $\mathbf{var}$, $\boldsymbol{p} - \boldsymbol{c}$, $\delta_i$, and features learned from RIFE stage where $\boldsymbol{m}$, $\mathbf{var}$ is the voxel statistics from $P_{\delta_i}$. After this per frame voxel feature augmentation, all the points are merged to one set $P$ followed by normal voxelization and point net. The rest of the model is the same as single frame models. 3) Given an input sequence $F = \{f_i | i = 0, 1, ...,\}$, frames are re-grouped into $\tilde{F} = \{(f_i, f_{i-1}, ..., f_{i-k}) | i = 0, 1, ...\}$ to train a $k + 1$-frame temporal RSN model with target output for frame $i$. If $i - k < 0$, we reuse the last valid frame.

## C. Ensemble Details

We provide additional description of the ensembling approach used to produce results highlighted in Table 3. We combine both data-level and test-time augmentation-based voting schemes: We trained five copies of the proposed model, each using a disjoint subset of 80% of the original training data. For each of the trained model, we perform box prediction under five random point cloud augmentations including random rotation and translation. This procedure yields 25 sets of results in total for each sample. We then use the box aggregation strategy proposed by Solovyev et al.[1], extended to 3D boxes with a yaw heading.

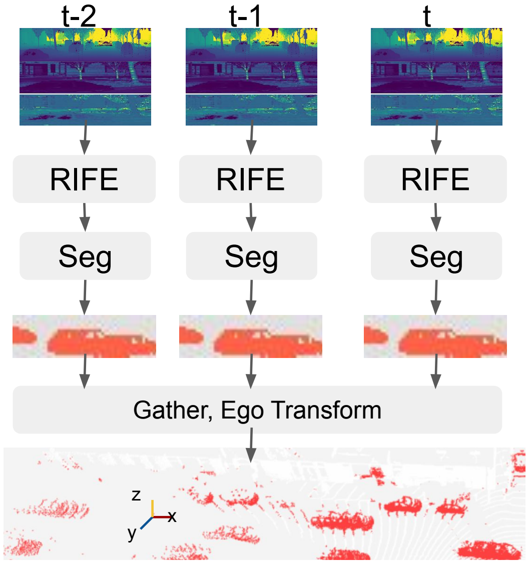

Figure 9. Expanded temporal RSN architecture before SPFE.

[1] Weighted Boxes Fusion: ensembling boxes for object detection models. Solovyev et al.