Exploiting Temporal Relations on Radar Perception for Autonomous Driving
Appendix

Peizhao Li[1]*, Pu Wang[2], Karl Berntorp[2], Hongfu Liu[1]
[1]Brandeis University, [2]Mitsubishi Electric Research Laboratories
{peizhaoli,hongfuliu}@brandeis.edu, {pwang,berntorp}@merl.com

A. Temproal Feature Extraction

We add Fig. 1 to illustrate the skip connections in the backbone neural networks. Skip connections within CNNs are designed to jointly involve high-level semantics and low-level finer details in output feature representation. Specially, we add three skip connections in ResNet and gradually up-sample the features from a deeper layer. The final feature representations are down-sampled with a ratio of 4 compared to the original inputs in this U-Net structure.

B. Multiple Object Tracking: Evaluation and the Decoding Algorithm

We adopt the series of MOT metrics [4] for evaluation. We pick several key metrics in experiments: MOTA (Multiple Object Tracking Accuracy), MOTP (Multiple Object Tracking Precision), ID switch (IDSW), track fragmentations (Frag.), mostly tracked (MT), and partially tracked (PT). The MOTA score is calculated by

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t},$$

where t is the frame index, GT is the number of ground-truth objects, FN and FP refer to false negative and false positive detection. The value of MOTA is in the range $(-\infty, 100]$. It can be deemed as the combination of detection and tracking performance, and is widely used as the main metric for accessing multiple object tracking quality. MOTP is the average IoU value on all ground-truth bounding boxes and its assigned prediction. It describes the localized precision. The rest of these metrics all reflect the quality of predicted tracklets. For detailed definitions and calculations of MOT metrics, please refer to [4].

We attach a decoding algorithm for multiple object tracking. The tracking algorithm mainly follows [9] which associates objects from successive frames purely based on the cost of Euclidean distance. The position of an object in the previous frame is complemented with a predictive positional tracking offset $\hat{\mathbf{d}}$ to infer its potential position in the next frame. Then, objects in previous and current frames are associated and propagate the object's ID in a bipartite graph with a greedy algorithm based on the distance between their center 2D positions. Empirically, we do not further extend a tracklet if it cannot find a matched candidate.

Algorithm 1 Multiple Object Tracking Decoding

Require: $T^{t-1} = \{(\mathbf{c}, \text{id})_j^{t-1}\}_{j=1}^{M}$: tracked objects in the previous frame $t-1$; $\hat{B}^t = \{(\hat{\mathbf{c}}, v, \hat{\mathbf{d}})_i^t\}_{i=1}^{N}$ heatmap predictions of object centers $\hat{\mathbf{c}}$, confidence v, and tracking offsets $\hat{\mathbf{d}}$. $\hat{B}^t$ are sorted in a descending order according to v. Distance threshold k. Birth threshold b.

$S \leftarrow \emptyset, T^t \leftarrow \emptyset$
$W \leftarrow \text{Cost}(\hat{B}^t, T^{t-1})$ ▷ $W_{ij} = ||\hat{\mathbf{c}}_i^t - \hat{\mathbf{d}}_i^t, \hat{\mathbf{c}}_j^{t-1}||_2$
for $i \leftarrow 1, N$ **do**
  $j \leftarrow \arg\min_{j \notin S} W_{ij}$
  if $w_{ij} \leq k$ **then**
    $T^t \leftarrow T^t \cup (\hat{\mathbf{c}}_i^t, \text{id}_j^{t-1})$ ▷ Propagate matched id
    $S \leftarrow S \cup \{j\}$ ▷ Mark candidate j as tracked
  else if $v_i \geq b$ **then**
    $T^t \leftarrow T^t \cup (\hat{\mathbf{c}}_i^t, \text{New id})$ ▷ Create a new track
  end if
end for
return T^t

C. Ablation Study

We add an experiment on split train good weather in Fig. 2 to analyze the change of the number of the selective feature vectors for temporal relational layers, where we vary the value K from 2 to 20. The detection performance consistently improved before K reached 8, but drop when continually increase the value of K. The scenario indicates involving redundant objects in relation modeling could slightly corrupt the temporal relation learning. The value of K should be selected based on the average number

*Work done during the internship at MERL

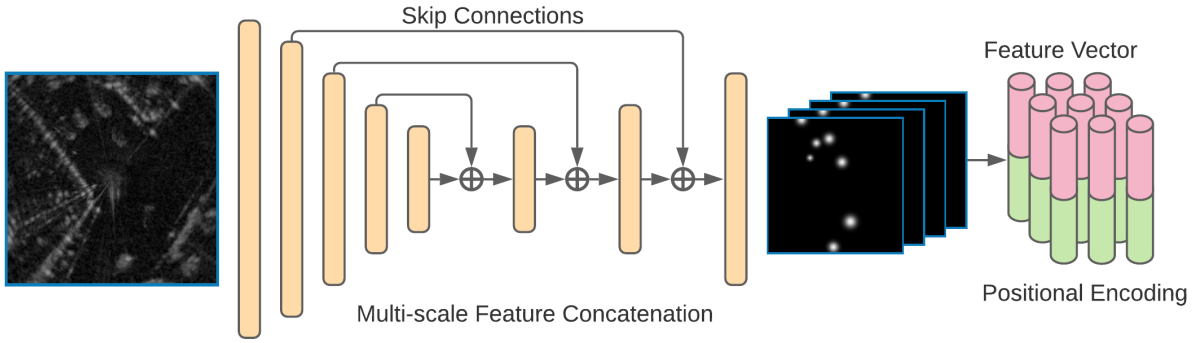

Figure 1. The backbone networks are inserted with several skip connections to collect features at different scales for predictions. Features selected for temporal relations modeling are attached with positional encoding to reveal the locality of objects.

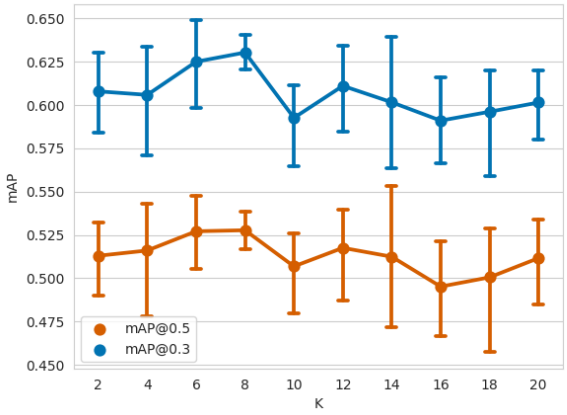

Figure 2. Detection performance with varying K value.

of objects per frame but not including excessive noise. We empirically set K to 8 in our experiments.

D. Additional Visualization Result

We present additional visualization results in Fig. 3 on object detection. In the detection, green bounding boxes are ground-truth annotations, while red are predictions. The same observations are confirmed in the additional visualizations. False positive predictions are mainly due to the 'ghost' objects in radar signals, and the rest are localized in the surroundings or outer space where the angular resolution is low.

E. A Short Review of Radar Dataset

Besides the algorithmic design, many radar datasets are emerging which are crucial for machine learning research. Among these datasets, radar data are currently presented in various data formats, *i.e.* radio frequency heatmap, radar reflection image, or point cloud. *RadarScenes* dataset [6] provide abundant point-wise annotations with doppler for automotive radar. However, there is no bounding box annotation for objects. *Carrada* dataset [5] records the range-angle and range-Doppler heatmap. Their data are mainly recorded in experimental sites like parking lots but not in real driving environment. *CRUW* dataset [8] offers radar's radio frequency images with camera-projected annotations. *nuScenes* [2] contains multi-modal data including Lidar, camera, and radar. However, radar data in *nuScenes* only afford sparse point cloud, while the Lidar and camera data are the main advantage of this dataset. *MulRan* [3] and *Oxford* [1] datasets present high-resolution radar images for urban driving scenarios but without object-level annotation. In our paper, we conduct detection and tracking experiments on point cloud-based radar images in adverse weather from *Radiate* dataset [7], and every significant object has bounding box and tracking ID annotations for training.

References

[1] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *2020 IEEE International Conference on Robotics and Automation*, pages 6433–6438, 2020. 2

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2

[3] Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. Mulran: Multimodal range dataset for urban place recognition. In *2020 IEEE International Conference on Robotics and Automation*, pages 6246–6253, 2020. 2

[4] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1