Supplementary Material — Learning Lane Graph Representations for Motion Forecasting

Ming Liang[1], Bin Yang[1,2], Rui Hu[1], Yun Chen[1],
Renjie Liao[1,2], Song Feng[1], and Raquel Urtasun[1,2]

[1] Uber ATG
[2] University of Toronto
{ming.liang,byang10,rui.hu,yun.chen,rjliao,songf,urtasun}@uber.com

Abstract. In this supplementary material we first show the detailed network architecture. We then showcase qualitative results. Additionally we include a video showcasing predictions of our approach.

1 Network Architecture

We show the detailed architecture of our model in Figure 1. Our model is composed of 4 modules, ActorNet, MapNet, FusionNet, and the Prediction Header. ActorNet extracts temporal features with a 1D CNN and merges the multi-scale features with a Feature Pyramid Network [3]. MapNet is a LaneGCN, which consists of a stack of 4 multi-scale LaneConv residual blocks (see Fig. 4 in the main paper) and extracts lane topology features.. FusionNet is a stack of 4 interaction modules, including actor-to-lane (A2L), lane-to-lane (L2L), lane-to-actor (L2A), actor-to-actor (A2A). Each of A2L, L2A and A2A consists of 2 attention residual blocks. L2L is another LaneGCN. Finally, the updated actor features are used by the prediction header to produce the multi-modal trajectories and their confidence scores.

2 Qualitative Results

We show more qualitative results on the Argoverse [1] dataset in Figure 2 and Figure 3. The Argoverse dataset provides the centroid locations of all objects in each sequence. Each sequence has an interesting object called **agent** whose future trajectory needs to be predicted. In Figure 2 and Figure 3, all trajectories end with a circle. Past agent trajectories are depicted in yellow, multi-modal agent predictions in green, ground truth agent trajectories in red, other actors' ground truth trajectories in blue.

3 Video

We also make a video depicting the predicted agent trajectory and the ground truth (see **video.mp4**). The experiment is conducted on the Argoverse dataset.

As indicated in Section 2, there is only one actor of interest (the agent) in each sequence. In the video, we show the predicted agent trajectory of our model (green circle, the best mode), ground truth agent trajectory (red circle), and ground truth trajectory of other actors (blue circle).

References

1. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8748–8757 (2019)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Lin, T.Y., Dollár, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 936–944 (2016)

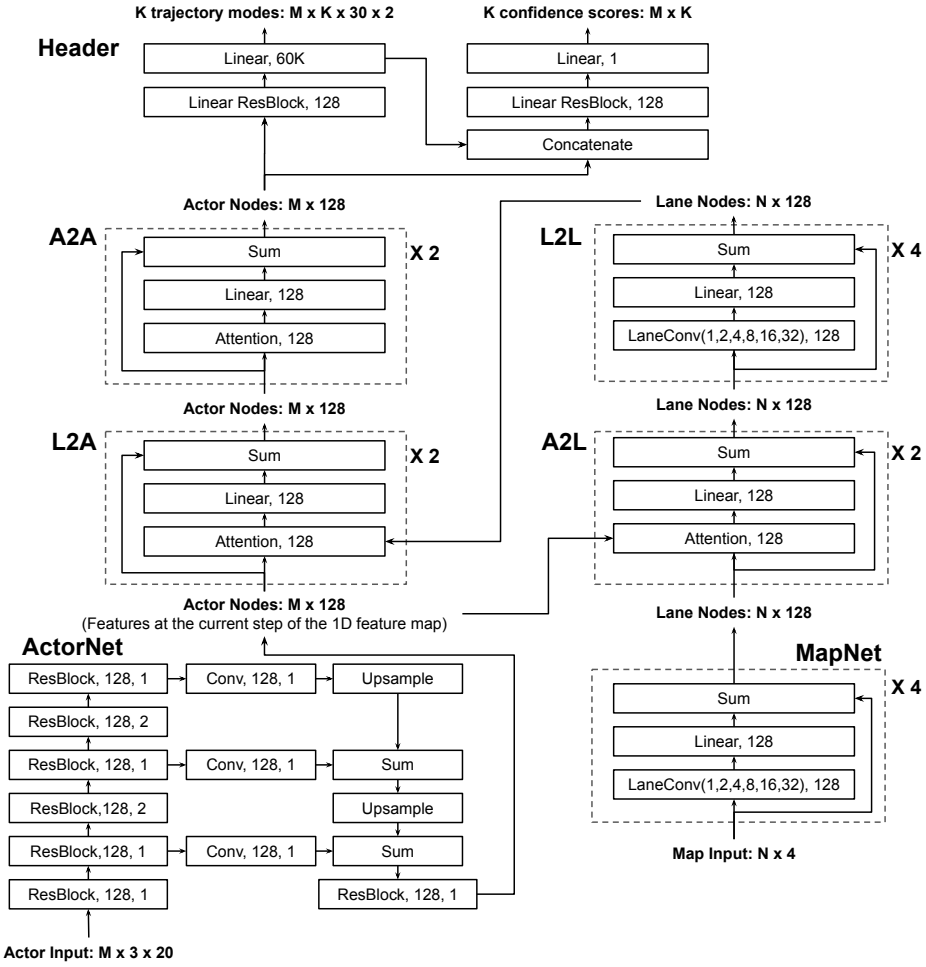

Fig. 1. Network architecture:. **ResBlock, 128, 1** denotes an 1D convolution residual block [2] with 128 output channels and stride 1. **Linear ResBlock, 128** denotes a linear residual block [2] with 128 output channels. **Conv, 128, 1** denotes an 1D convolution layer with 128 output channels and stride 1. **Linear, 128** denotes a linear layer with 128 output channels. **Upsample** denotes a bilinear upsampling layer with scale factor 2. **Sum** denotes an element-wise summation layer. **Concatenate** denotes a concatenation layer along the feature channel dimension. **LaneConv(1,2,4,8,16,32), 128** denotes a multi-scale LaneConv layer with 128 output channels (see Equation (4) in the main paper). **Attention, 128** denotes an attention layer with 128 output channels (see Equation (5) in the main paper).

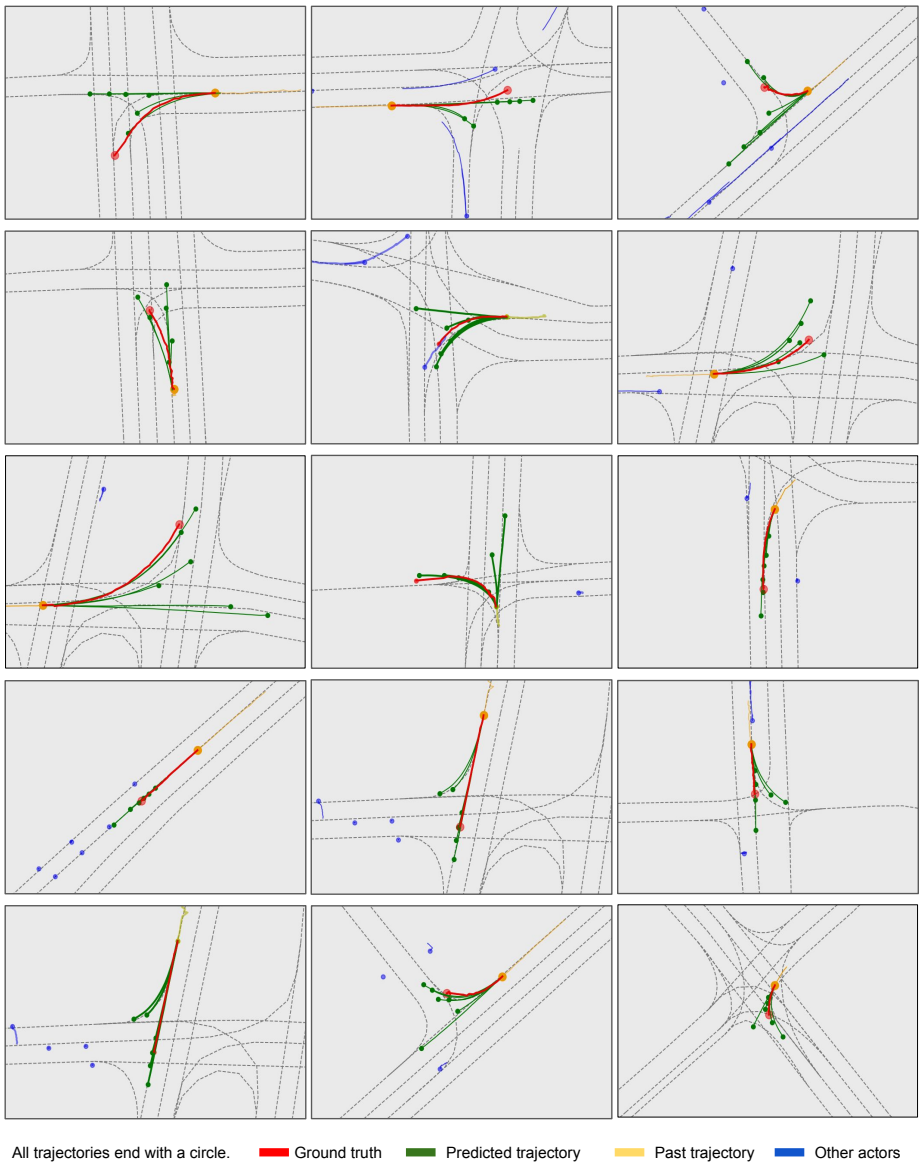

Fig. 2. Qualitative Results.

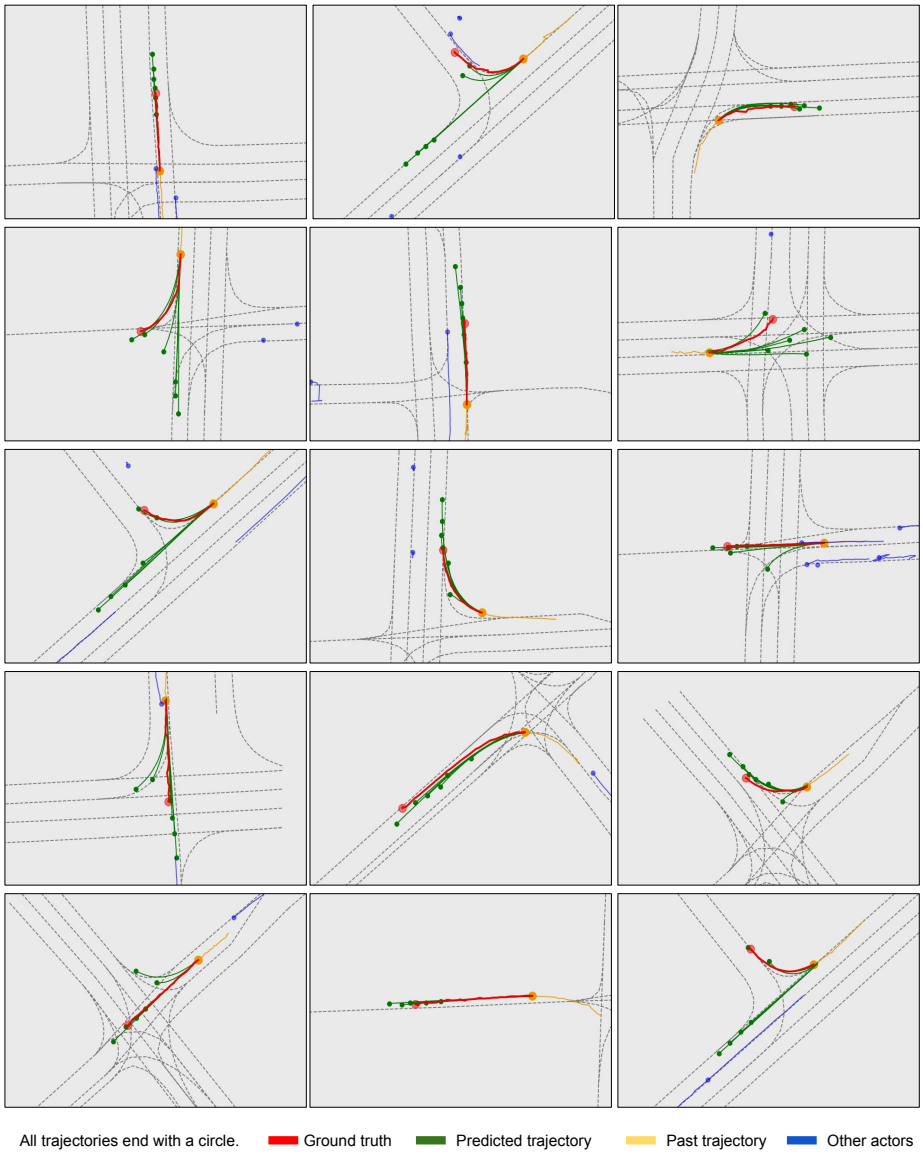

Fig. 3. Qualitative Results.