CCLSTM: Coupled Convolutional Long-Short Term Memory Network for Occupancy Flow Forecasting

Supplementary Material

A. Ablation Study on Receptive field

We conduct an ablation study to assess how spatial downsampling ratios, which balance spatial resolution and receptive field size, affect model performance. To ensure a fair comparison across architectures, we keep the number of parameters and final embedding dimensionality consistent across all encoder–decoder configurations (Tab. 5). We vary the spatial scale from 1/4 to 1/32 and using both 3×3 and 5×5 kernels in the forecasting module. The results show that a downsampling scale of 1/8 yields the best overall performance under the given constraints. Architectural details are illustrated in Fig. 9, and the corresponding performance results are presented in Fig. 10.

Spatial Scale	**Convolutional Layer Output Size**					
	1	2	3	4	5	6
1/4	64	128	256			
1/8	32	64	128	256		
1/16	16	32	64	128	256	
1/32	8	16	32	64	128	256

Table 5. **Encoder/Decoder Architecture:** Network configuration parameters ensuring consistent parameter count and final embedding dimensionality across different spatial scales.

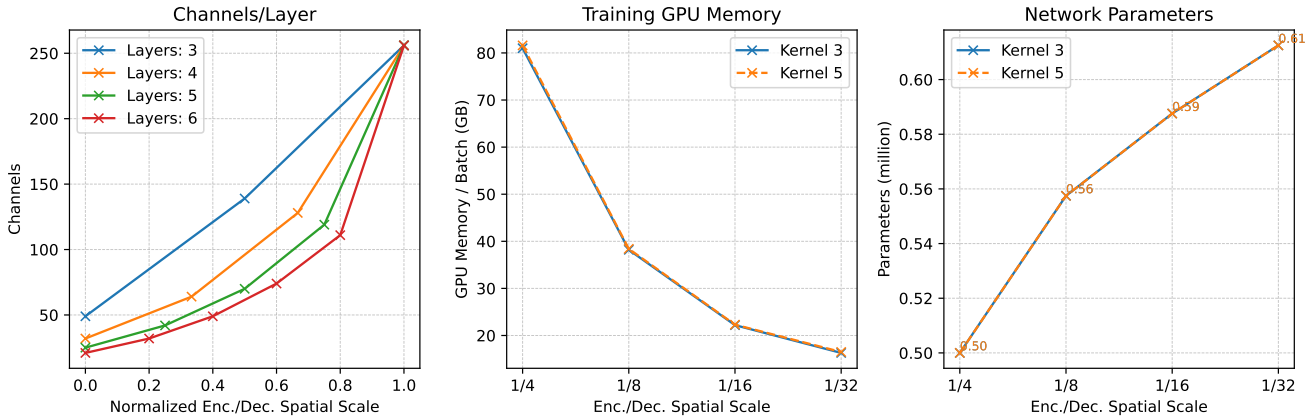

Figure 9. **Network architecture characteristics:** (1) Output and input channel dimensions for the encoder and decoder, respectively; (2) GPU memory requirements for training with a batch size of 6 across different spatial scales; (3) number of encoder parameters for each spatial scale.

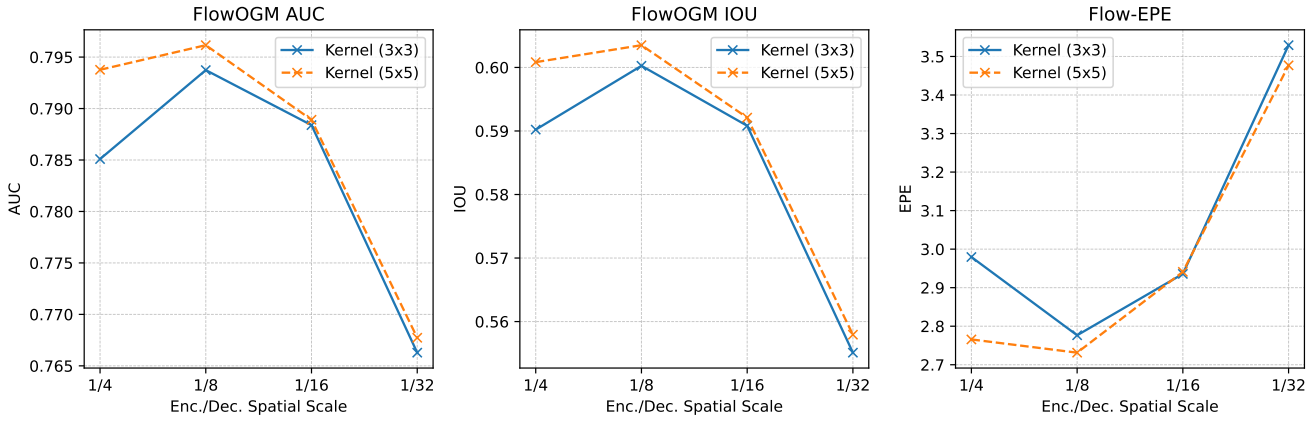

Figure 10. **Metrics results:** Comparison of different spatial scales for occupancy and flow prediction on the test set. Models were trained using the original setup, except with the batch size reduced to 6.

B. Additional Qualitative Results

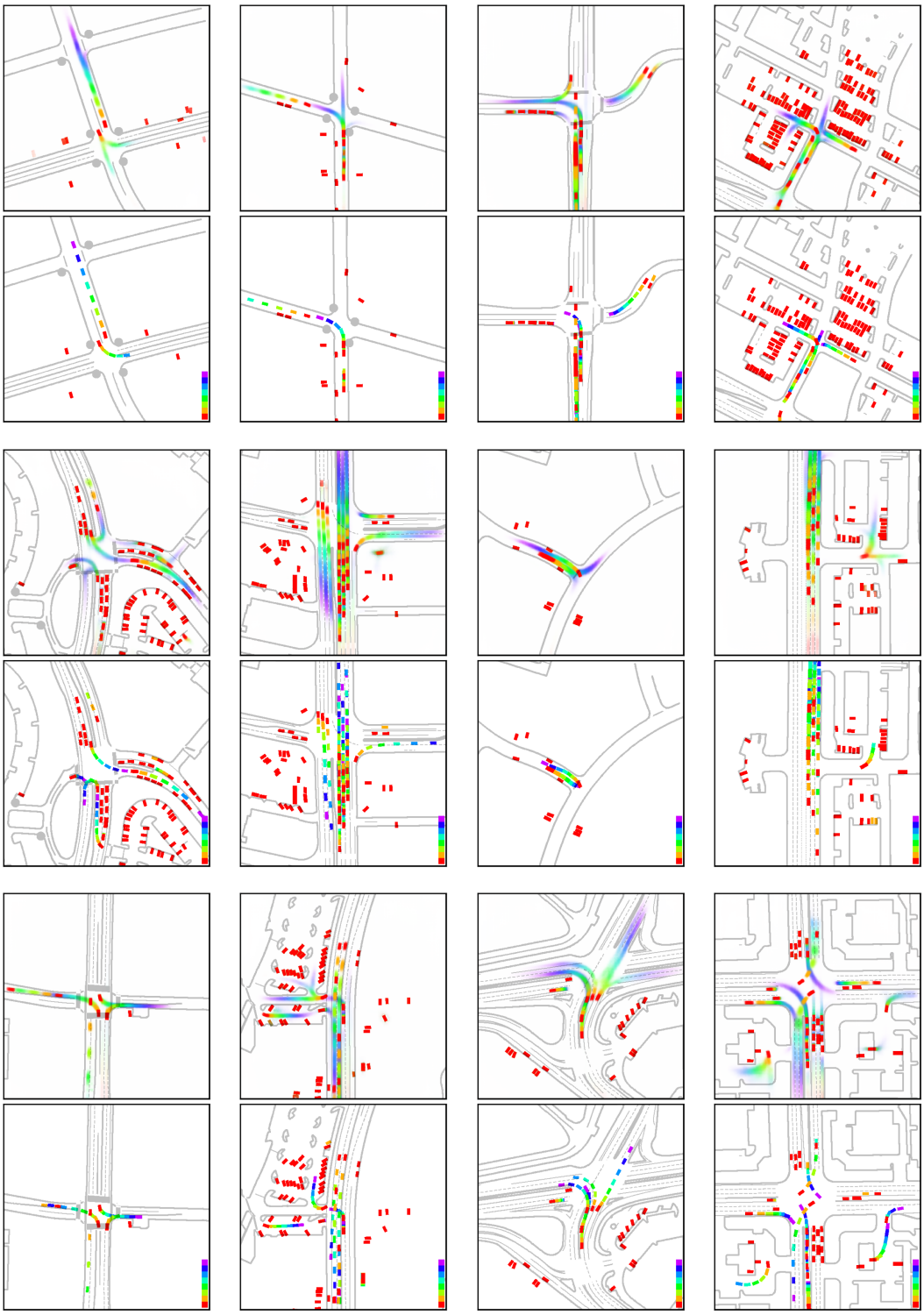

Figure 11. **Qualitative results on WOMD validation set**. Each subplot displays the testing result of (1) color coded future occupancy prediction, (2) color coded future occupancy target. The results are the outputs of our state-of-the-art model using $H, W = 512$ input rasters. Color coding denotes timesteps $t \in [1, T_f]$ with $red = 1$.

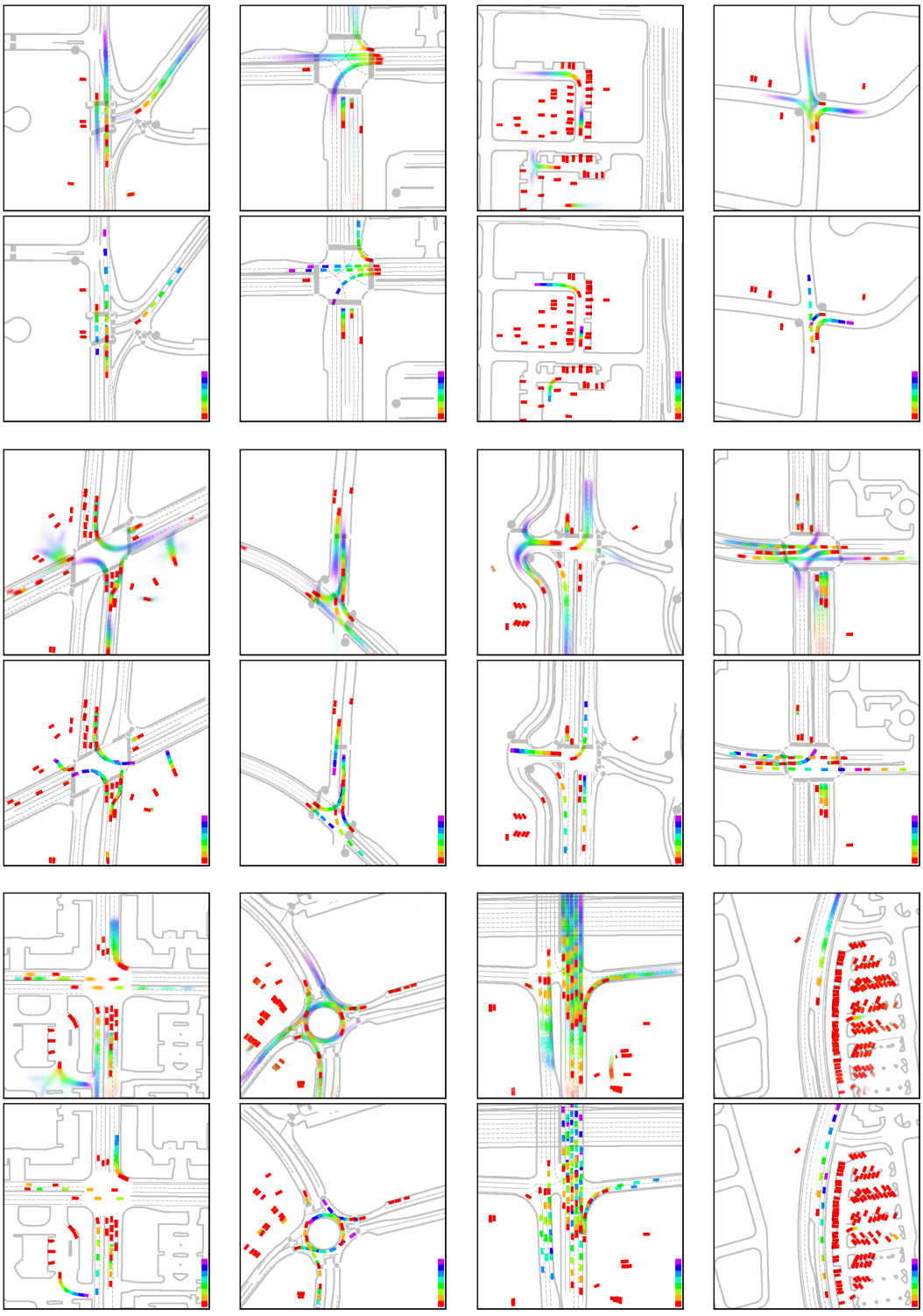

Figure 12. **Qualitative results on WOMD validation set**. Each subplot displays the testing result of (1) color coded future occupancy prediction, (2) color coded future occupancy target. The results are the outputs of our state-of-the-art model using $H, W = 512$ input rasters. Color coding denotes timesteps $t \in [1, T_f]$ with $red = 1$.