Supplementary Material

A. Impact of Network Structure

We have used a 3DCNN-based network in the main body for the primary task. One interesting question is whether the same improvement will be achieved if we use a different network for the primary task. To answer this question, we construct an alternative network architecture for future depth prediction based on convolutional LSTM (ConvLSTM) [3]. ConvLSTM has been commonly used to extract spatio-temporal information from videos [2].

We follow [4] to construct a ConvLSTM network for the primary task. The original ConvLSTM in [4] aims to predict the depth of the current frame. We thus use a modified structure for future depth prediction. Specifically, we do not predict the current depth at each time step. Instead, we directly output the future depth from the last hidden unit with a convolutional layer. To introduce the self-supervised auxiliary task into the network, we then symmetrically place a single convolutional layer after the last hidden output for frame reconstruction. The detailed architecture of the ConvLSTM network is provided in the supplementary material. The performance of using ConvLSTM instead 3DCNN in our framework is shown in Table 1. We observe a similar trend in terms of performance. Our proposed method significantly outperforms our alternatives. The experimental results provide convincing evidence that our meta-auxiliary learning approach is agnostic to the backbone architecture.

Method	Error (lower is better)				Accuracy (higher is better)		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Impact of network architecture							
ConvLSTM (Primary only)	0.110	0.697	4.126	0.167	0.883	0.970	0.989
ConvLSTM (Multi-task)	0.109	0.657	4.112	0.165	0.884	0.970	0.989
ConvLSTM (Ours)	**0.107**	**0.648**	**4.108**	**0.163**	**0.886**	**0.971**	**0.990**

Table 1: Ablation studies of LSTM model: the performance of our method when using ConvLSTM instead of 3DCNN as the model architecture. We observe that the performance follows the same trend.

Figure 1 illustrates the detailed structure of our ConvLSTM. Since we have four input frames $(I_{t-3}, I_{t-2}, I_{t-1}, I_t)$, the ConvSLTM module contains four time steps. We demonstrate a single ConvLSTM cell at time step t in Eq. 1.

$$
\begin{aligned}
F_t &= \delta([f_t, \mathcal{H}_{t-1}] * W_F + b_F), \\
i_t &= \delta([f_t, \mathcal{H}_{t-1}], *W_i + b_i), \\
C_t &= i_t * \mathit{tanh}([f_t, \mathcal{H}_{t-1}], *W_C + b_C) + C_{t-1} * F_t, \\
o_t &= \delta([f_t, \mathcal{H}_{t-1}] * W_o + b_o), \\
\mathcal{H}_t &= \mathit{Conv}([o_t, \mathit{tanh}(C_t)])
\end{aligned}
\tag{1}
$$

where f_t denotes the current feature map extracted from the spatial feature extraction network, * represents convolution operation. W and b term the kernels and bias of each convolutional layer. Note that the *Conv* contains two convolutional layers. To predict the desired future depth map at $t+1$, we take the outputs of the last layer of *Conv* at the last hidden state. Then a depth prediction layer is added to map the judicious high dimensional features to depth maps. Furthermore, we place a single 3×3 convolutional layer with ReLU [1] symmetrically to the depth prediction layer. This layer aims to reconstruct four RGB frames in one shot.

Similar to our proposed 3DCNN with auxiliary learning, the feature extractor and ConvLSTM can finally produce desired features for image reconstruction and future depth prediction.

B. Additional Quantitative Results

In this section, we additionally show some qualitative results for the ConvLSTM-based method and 3DCNN-based method.

Figure 2 shows the qualitative comparison of our proposed method with several baselines. It can be observed that the depth maps produced with meta-auxiliary learning are comparatively the best in terms of visual quality and accuracy. For example,

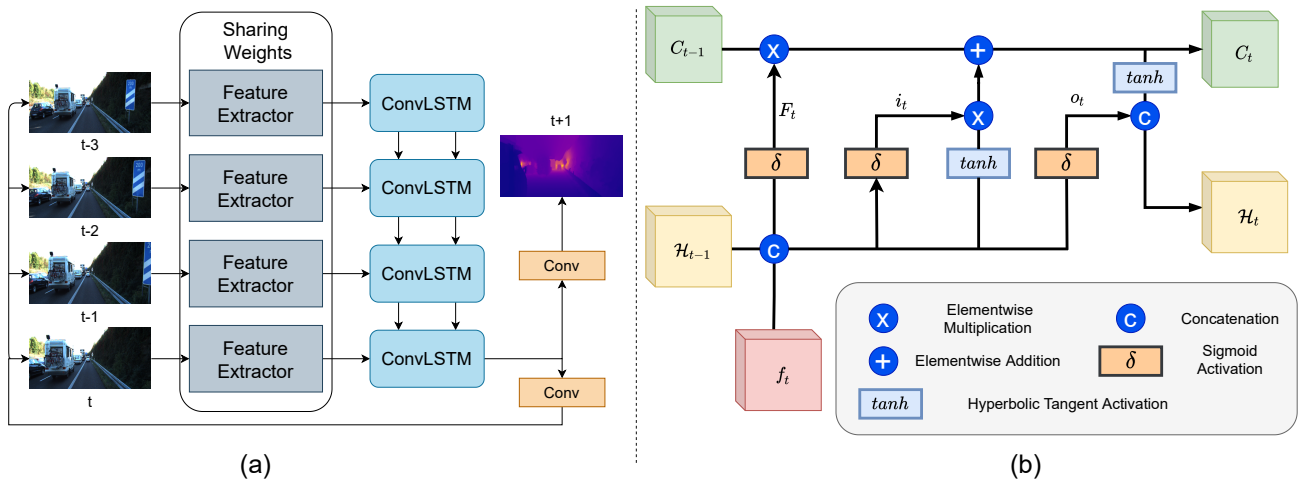

Figure 1: Illustration of our ConvLSTM architecture. (a) demonstrates the overall architecture of the ConvLSTM network for meta-auxiliary learning. (b) shows a single ConvLSTM cell in (a).

the edges of the traffic sign and the car are clearer, and objects are detected with a lower failure rate. Besides, colorized depth map also shows that our method can predict object distance with higher accuracy (*e.g.* the color on our predicted traffic sign is closest to the ground truth.).

Figure 3 further proves the effectiveness of our meta-auxiliary learned 3DCNN network with additional qualitative examples. By comparing our generated future depth maps with their correspondingly ground truth, we can find that our method is able to produce high accuracy future depth.

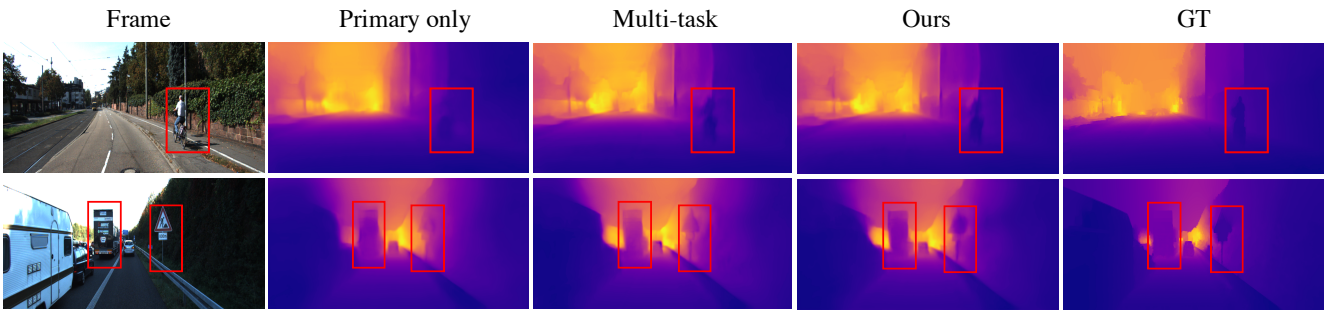

Figure 2: Quantitative results of ConvLSTM-based method. Primary only and multi-task denote the two baseline methods mentioned in our paper. Ours represent the meta-auxiliary learned ConvLSTM network. The figure illustrates that our method can predict depth maps with better object boundaries and object distance.

References

[1] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

[2] Seyed shahabeddin Nabavi, Mrigank Rochan, and Yang Wang. Future semantic segmentation with convolutional lstm. In *British Machine Vision Conference*, 2018.

[3] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 2015.

[4] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *IEEE/CVF International Conference on Computer Vision*, pages 1725–1734, 2019.

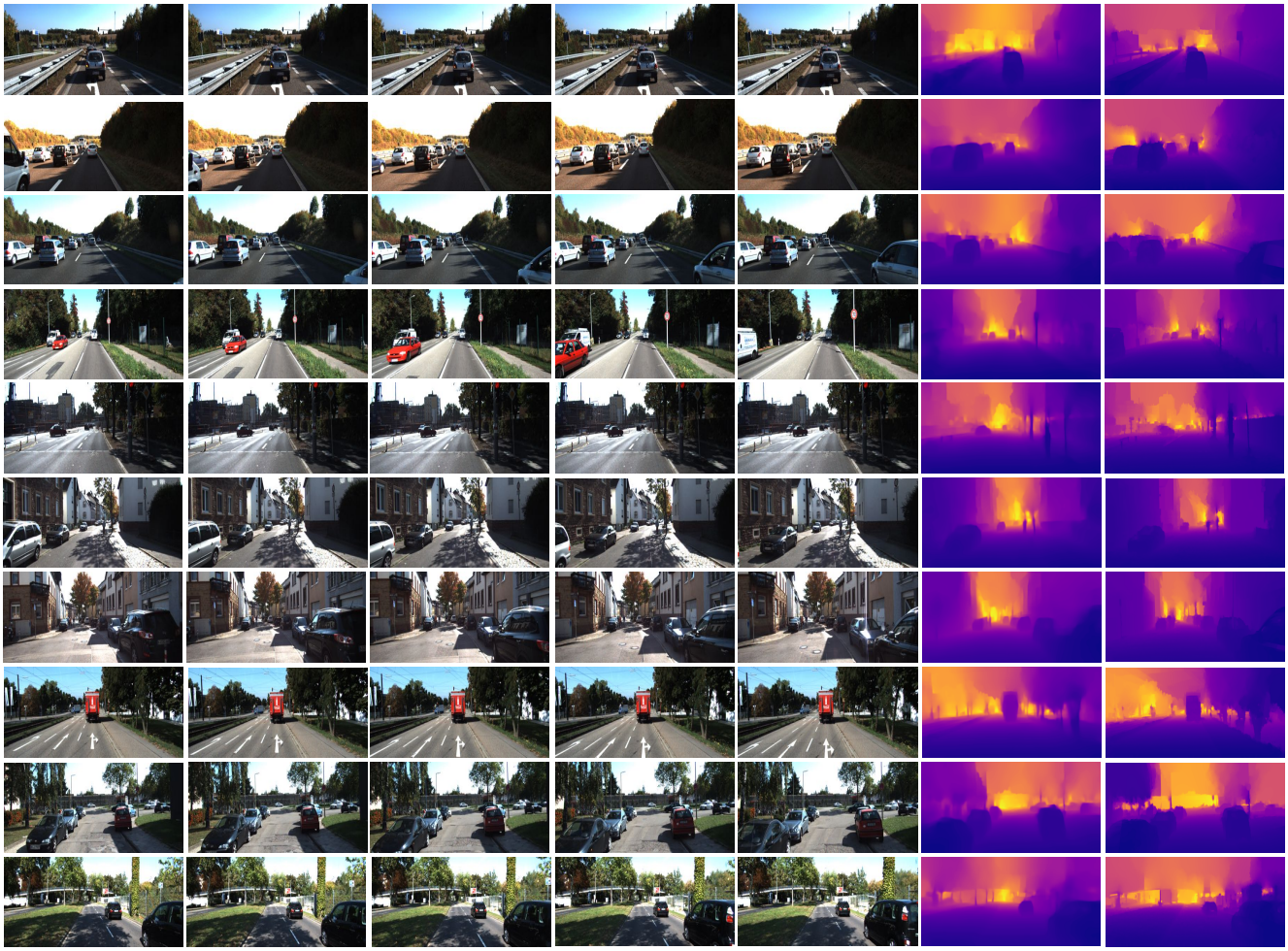

Figure 3: Additional quantitative results of the 3DCNN-based method. We show the input video sequences with t-3, t-2, t-1 and t. t+1 denotes the future RGB frame, which is not touched during the depth prediction. Ours represents the depth map predicted from our meta-auxiliary learned 3DCNN network, and GT is the ground truth future depth map.