Supplementary Materials for Less is More: Consistent Video Depth Estimation with Masked Frames Modeling

This document involves the following contents:

- Mask sampling strategies.
- Masking ratios for inference.
- Depth estimation metrics.
- Comparison with structure-from-motion methods.
- Comparison with single image depth estimation methods.
- Ablation of different backbones.
- Qualitative depth results.
- Network architecture of our depth predictor.

1 MASK SAMPLING STRATEGIES

In this section, we illustrate our ablation study on the mask sampling strategies. We adopt the random masking for training in our approach. In Table 1, we also train our FMNet in the uniform manner, which means we use fixed and uniform masking during training. For example, we will retain the fourth and eighth frames with twelve frames input. We keep the same masking ratio of 83.33% for the comparison of sampling strategies.

In Table 1, we can see that our random masking strategy achieves higher depth accuracy and better temporal consistency. The random masking can be considered as a form of data argumentation. In this way, our FMNet can learn temporal correlations with various time intervals, while the uniform masking strategy can only model correlations of a fixed length of time such as 4 frames. As a consequence, we adopt the random masking strategy for training.

Table 1: Ablation study on mask sampling strategies. We keep the same masking ratio of 83.33% for the comparison of sampling strategies. In order to reduce the experimental cost, we randomly choose 40 videos for training and 10 videos for OPW evaluation on the NYU Depth V2 dataset [12] in this experiment. The depth metrics are still evaluated on the public test dataset with 654 samples, which can not be compared with the results of our FMNet on the full dataset. The way of random masking achieves higher spatial accuracy and better temporal consistency. Best performance is in boldface.

Sampling	Rel	RMSE	log 10	δ_1	δ_2	δ_3	*OPW*
Uniform	0.253	**0.725**	0.095	0.622	0.875	0.965	8.063
Random	**0.221**	0.738	**0.093**	**0.628**	**0.889**	**0.968**	**6.965**

2 MASKING RATIOS FOR INFERENCE

We also ablate our masking ratios for inference in Fig. 1. In our approach, we use uniform masking for inference to avoid randomness in our depth prediction results. For example, with twelve frames input, 83.33% means that we mask ten frames and retain the fourth and eighth frames. In this experiment, we use the same model trained with 83.33% random masking on 40 videos in Sec. 1. The *OPW* is evaluated on the same 10 videos.

We can see that inferring with lower masking ratios causes a decrease of consistency due to higher redundancy. We also try the extreme situation: inference without masking. We directly feed input sequences without masking to our temporal structure encoder. In this way, our FMNet loses the vital mechanism of masked frames predicting. Reconstructing masked frames according to the unmasked ones plays a significant role in temporal consistency.

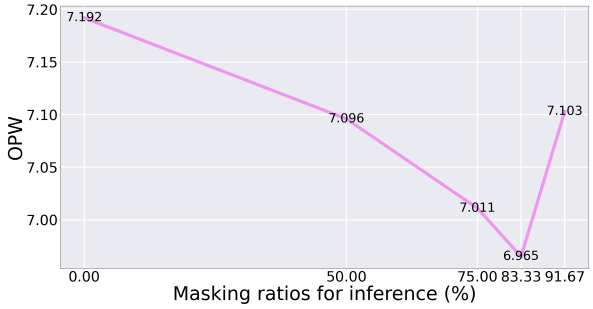

Figure 1: Ablation study on masking ratios for inference. The X-axis represents masking ratios and the Y-axis means OPW. Here we use the same random masking model in Table 1.

In Fig. 2, based on our FMNet trained on the full NYU depth V2 dataset [12], we compare the visual depth results of 83.33% and 50% masking ratios for inference. The qualitative results of 50% masking ratios have worse consistency and flickering than 83.33% masking ratios due to higher temporal redundancy.

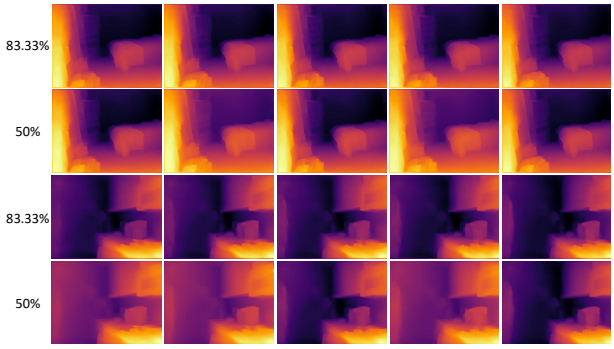

Figure 2: Visual results comparison of different masking ratios for inference. The results are produced by our FMNet on the full NYU Depth V2 dataset [12], which is trained with the random masking ratio of 83.33%.

3 DEPTH ESTIMATION METRICS

We adopt the commonly applied depth estimation metrics defined as follows:

- Mean relative error (REL): $\frac{1}{n}\sum_{i=1}^{n}\frac{||d_i-d_i^*||_1}{d_i^*}$;
- Root mean squared error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i-d_i^*)^2}$;
- Mean $\log_{10}$ error (log 10): $\frac{1}{n}\sum_{i=1}^{n}||\log_{10}d_i-\log_{10}d_i^*||_1$;
- Accuracy with threshold t: Percentage of d_i such that $max(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i})=\delta<t\in\left[1.25,1.25^2,1.25^3\right]$,

where n denotes the total number of pixels, d_i and d_i^* are estimated and ground truth depth of pixel i, respectively.

4 COMPARISON WITH STRUCTURE-FROM-MOTION METHODS

We show the comparison with structure-from-motion (SFM) methods on the KITTI dataset in Table 2. The quantitative results of structure-from-motion based methods (e.g., DeepV2D [14]) seem higher than the methods on the first four rows. However, those two kinds of methods are in different settings.

Structure-from-motion methods predict depth maps by feature matching over multiple frames. According to CVD [8], this idea benefits static scenes but brings an unavoidable defect which is that these methods "do not account for dynamically moving objects". They heavily rely on explicit motion segmentation. For example, they need to mask the moving cars or people for SFM and pose estimation. When their methods are used for videos with natural scenes or obvious objects motion, those methods inevitably fail. By contrast, our method is not limited by SFM and pose estimation.

In conclusion, SFM-based methods fit the bias of KITTI dataset, hence, previous works, such as dynamic-video-depth [20] and Cao *et al.* [2], exclude the structure-from-motion methods in their comparison list. We just follow the same setting and add some latest works to our comparison such as CVD [8] and Cao *et al.* [2].

Table 2: Comparison with structure-from-motion methods on the KITTI dataset. The structure-from-motion methods are on the last three rows and other methods are on the first four rows.

Method	Rel	RMSE	log 10	δ_1	δ_2	δ_3
ST-CLSTM [18] (ICCV 2019)	0.101	4.137	0.043	0.890	0.970	0.989
CVD [8] (ACM SIGGRAPH 2020)	0.130	4.876	–	0.878	0.946	0.970
Cao *et al.* [2] (ACM MM 2021)	0.109	4.366	0.047	0.872	0.962	0.986
Ours	0.099	3.832	0.042	0.886	0.968	0.989
BA-Net [13] (ICLR 2018)	0.083	3.640	–	–	–	–
DeepV2D [14] (ICLR 2020)	0.037	2.005	–	0.977	0.993	0.997
Wang *et al.* [15] (CVPR 2021)	0.034	1.919	–	0.989	0.998	0.999

5 COMPARISON WITH SINGLE IMAGE DEPTH ESTIMATION METHODS

Single image depth estimation methods [1, 5, 10, 11, 17] only take spatial depth accuracy into account and totally ignore the temporal depth consistency. As shown in Table 3, these methods achieve better performance in terms of spatial metrics, however, they suffer from obvious temporal inconsistency on video data. By contrast, consistent video depth estimation methods achieve much better temporal consistency. The core task of consistent video depth estimation is to remove flickering in video depth results. SC-GAN [16] seems to train their model on video data, however, their motivation and proposed solution only lie in spatial accuracy. This shows that these two types of methods are under two different settings. One is trying to achieve higher depth accuracy but totally ignoring the consistency; the other is trying to achieve consistent depth estimation of videos with good depth accuracy. In some real-world applications, e.g., 2D-to-3D video conversion [4] and video bokeh rendering [9, 19], depth consistency plays a vital role. Weird and obvious artifacts can be found if video depth is inconsistent.

Table 3: Comparison with single image depth estimation methods on the KITTI dataset. The first four rows are consistent video depth methods. The last five rows are methods only for spatial depth accuracy.

Method	Rel	RMSE	log 10	δ_1	δ_2	δ_3	*OPW*
ST-CLSTM [18] (ICCV 2019)	0.101	4.137	0.043	0.890	0.970	0.989	–
CVD [8] (ACM SIGGRAPH 2020)	0.130	4.876	–	0.878	0.946	0.970	34.741
Cao *et al.* [2] (ACM MM 2021)	0.109	4.366	0.047	0.872	0.962	0.986	–
Ours	0.099	3.832	0.042	0.886	0.968	0.989	30.596
VNL [17] (ICCV 2019)	0.072	3.258	–	0.938	0.990	0.998	45.295
BTS [5]	0.056	1.925	–	0.964	0.994	0.999	44.583
DPT [10] (ICCV 2021)	0.062	2.573	–	0.959	0.995	0.999	43.207
SC-GAN [16] (ICCV 2019)	0.063	2.129	–	0.961	0.993	0.998	–
AdaBins [1] (CVPR 2021)	0.058	2.360	–	0.964	0.995	0.999	43.841

Meanwhile, the training datasets and testing protocols are quite different between these two kinds of methods. For example, DPT [10], which is one of the state-of-the-art models for single image depth estimation, trains on 1.4 million images. Midas [11] is also based on mixing data from five different datasets. However, most of those datasets only contain single images. There is no such large scale public video depth dataset for now. Besides, some testing protocols are different. For example, Midas and DPT conduct scale and shift alignments for each testing image, while video depth methods such as ST-CLSTM [18], Cao *et al.* [2], and our methods do not.

As a consequence, previous works (ST-CLSTM [18], CVD [8], and Cao *et al.* [2]) exclude the single-image depth estimation methods in their comparison lists. We just follow the setting and add some latest works to our comparisons such as CVD [8] and Cao *et al.* [2].

6 ABLATION OF DIFFERENT BACKBONES

We also conduct ablation study of different backbones on the KITTI dataset. The results are shown in Table 4. Our FMNet can be easily extended to different backbones (the spatial structure feature extractor), which demonstrate the generality of our proposed method. Our FMNet achieves better performance than the model of Cao *et al.* [2] with the same backbone.

7 QUALITATIVE DEPTH RESULTS

In this section, we show additional qualitative depth results on the NYU Depth V2 dataset [12] in Fig. 3, Fig. 4, and Fig. 5. Visual results on the KITTI dataset [3] are in Fig. 6, Fig. 7, and Fig. 8. We compare the results of ST-CLSTM [18], our baseline, and our FMNet. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the

Table 4: Ablation study of different backbones on the KITTI dataset. The first four rows are our methods. The last two rows are results of Cao *et al.* [2].

Method	Backbone	Rel	RMSE	log 10	δ_1	δ_2	δ_3
Ours	ResNet18	0.105	3.936	0.045	0.875	0.965	0.988
Ours	ResNet50	0.105	3.893	0.044	0.876	0.965	0.988
Ours	ResNet101	0.101	3.868	0.043	0.882	0.967	0.989
Ours	ResNext101	0.099	3.828	0.042	0.886	0.968	0.989
Cao *et al.* [2] (ACM MM 2021)	ResNet18	0.109	4.366	0.047	0.872	0.962	0.986
Cao *et al.* [2] (ACM MM 2021)	ResNet101	0.106	4.243	0.045	0.879	0.964	0.986

last column. Our FMNet shows higher spatial accuracy and better temporal consistency. Please also refer to the supplementary video for more video depth visualization results.

8 NETWORK ARCHITECTURE OF OUR DEPTH PREDICTOR

The architecture of our depth predictor is illustrated in Fig. 9. The depth predictor contains five up-projection modules to gradually improve the spatial resolution and decrease the number of channels. In order to fuse the spatial and temporal structure features, we use the feature fusion module (FFM) [6, 7] and skip connection from the spatial structure feature extractor to the depth predictor. The temporal structure features could improve the inter-frames temporal consistency and the spatial features could help to reconstruct the detailed information in our depth results. At last, an adaptive output module is used to adjust the channel numbers and restore the final depth results.

REFERENCES

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4009–4018.

[2] Yuanzhouhan Cao, Yidong Li, Haokui Zhang, Chao Ren, and Yifan Liu. 2021. Learning Structure Affinity for Video Depth Estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 190–198.

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[4] Kevin Karsch, Ce Liu, and Sing Bing Kang. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence* 36, 11 (2014), 2144–2158.

[5] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326* (2019).

[6] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1925–1934.

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[8] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. 2020. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)* 39, 4 (2020), 71–1.

[9] Juewen Peng, Zhiguo Cao, Xianrui Luo, Hao Lu, Ke Xian, and Jianming Zhang. 2022. BokehMe: When Neural Rendering Meets Classical Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16283–16292.

[10] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12179–12188.

[11] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 03 (2020), 1623–1637.

[12] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*. Springer, 746–760.

[13] Chengzhou Tang and Ping Tan. 2018. BA-Net: Dense Bundle Adjustment Networks. In *International Conference on Learning Representations*.

[14] Zachary Teed and Jia Deng. 2019. DeepV2D: Video to Depth with Differentiable Structure from Motion. In *International Conference on Learning Representations*.

[15] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. 2021. Deep two-view structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8953–8962.

[16] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. 2019. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7494–7504.

[17] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. 2019. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5684–5693.

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

[19] Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. 2019. Synthetic defocus and look-ahead autofocus for casual videography. *ACM Transactions on Graphics (TOG)* 38, 4 (2019).

[20] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. 2021. Consistent depth of moving objects in video. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–12.

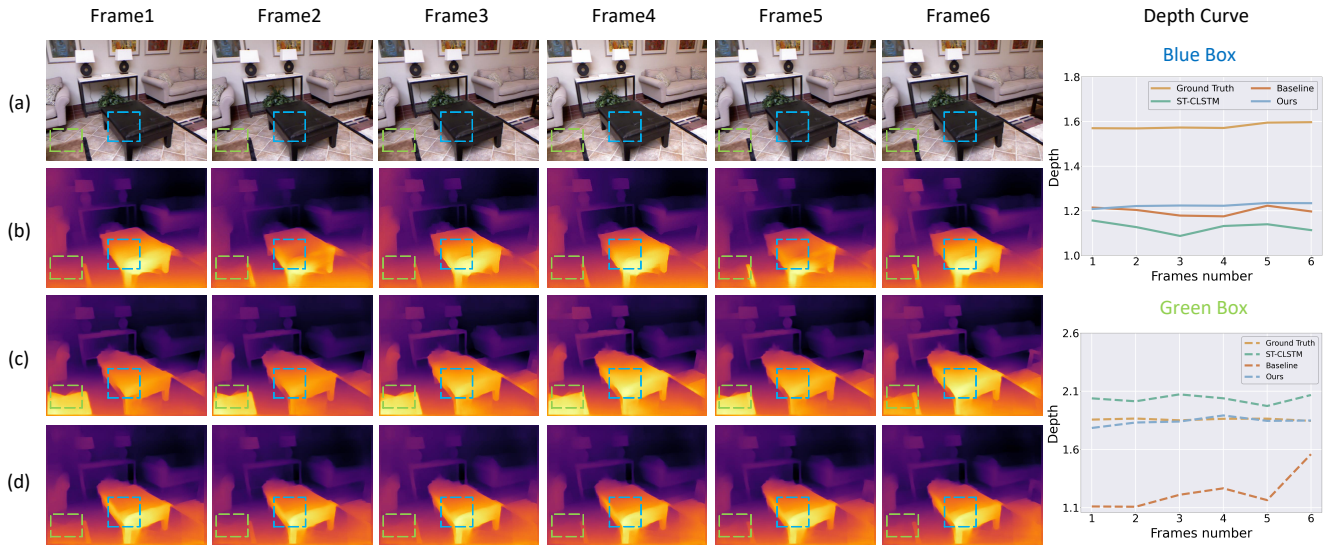

Figure 3: Qualitative depth results on the NYU Depth V2 dataset [12]. The four rows are: (a) RGB inputs; (b) Results of ST-CLSTM [18]; (c) Baseline results; (d) Results of our FMNet. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

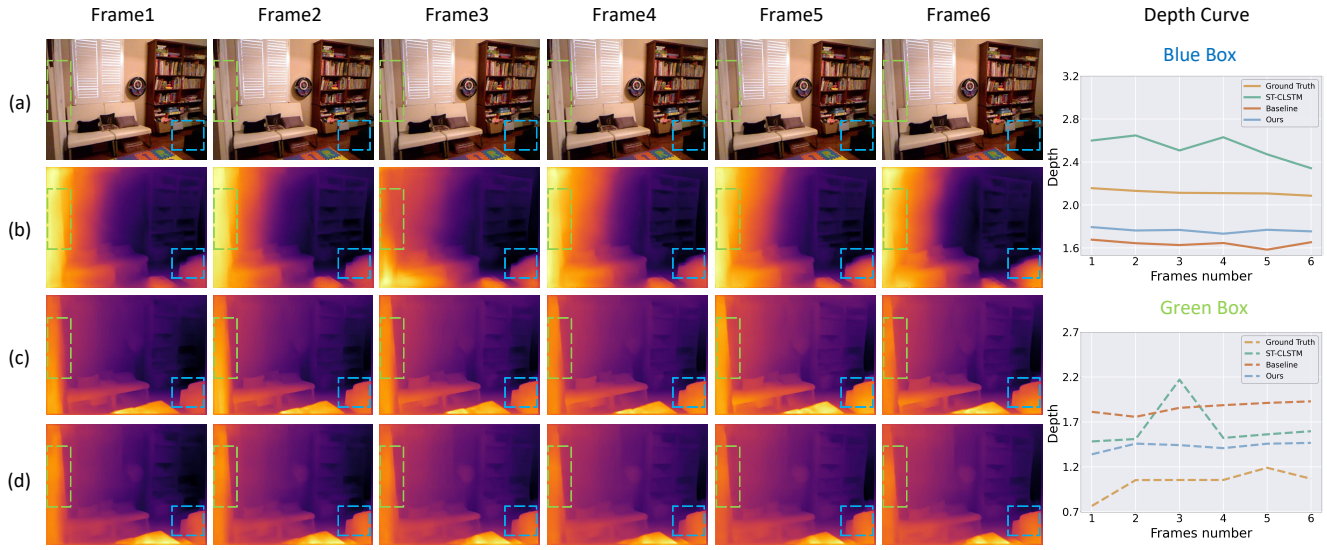

Figure 4: Qualitative depth results on the NYU Depth V2 dataset [12]. The four rows are: (a) RGB inputs; (b) Results of ST-CLSTM [18]; (c) Baseline results; (d) Results of our FMNet. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

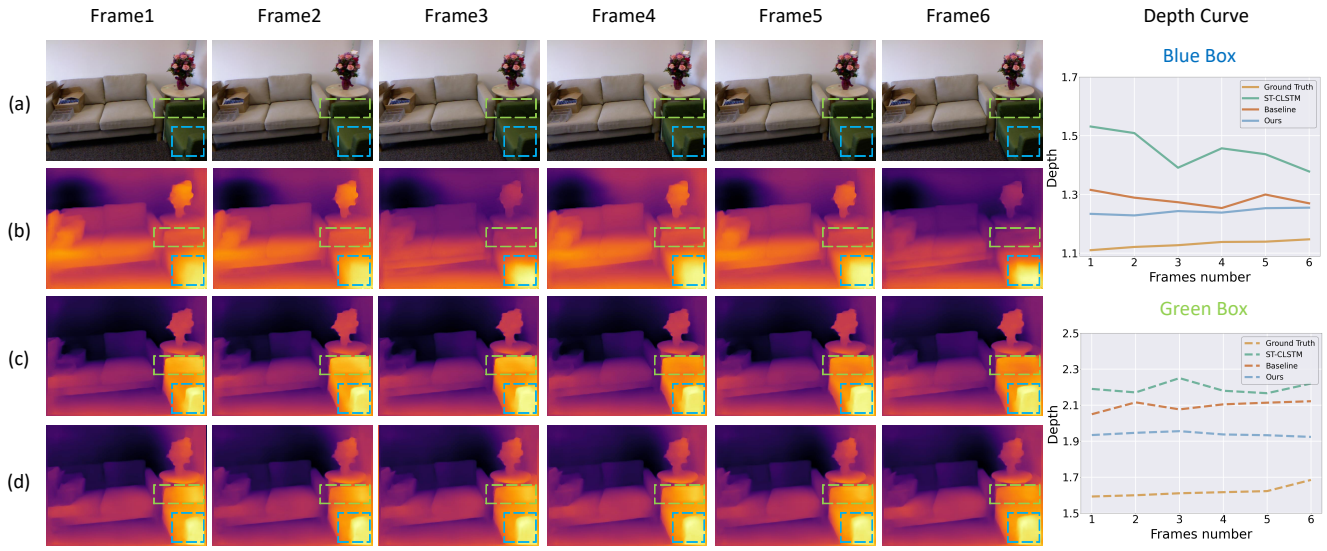

Figure 5: Qualitative depth results on the NYU Depth V2 dataset [12]. The four rows are: (a) RGB inputs; (b) Results of ST-CLSTM [18]; (c) Baseline results; (d) Results of our FMNet. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

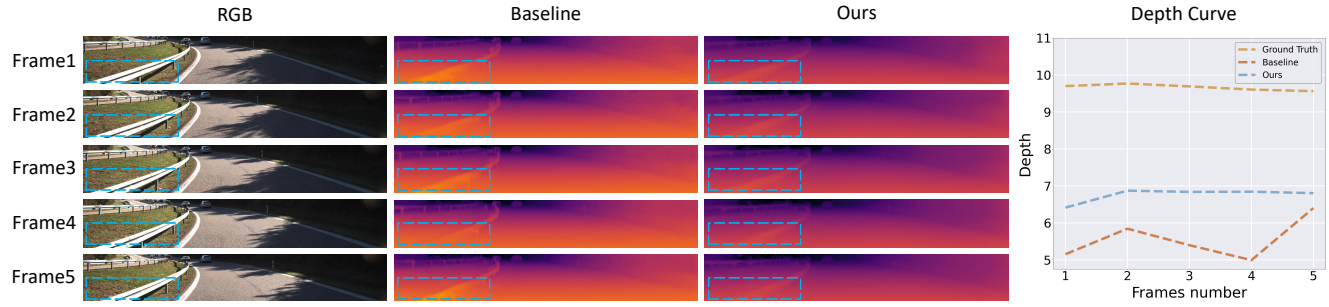

Figure 6: Qualitative depth results on the KITTI dataset [3]. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

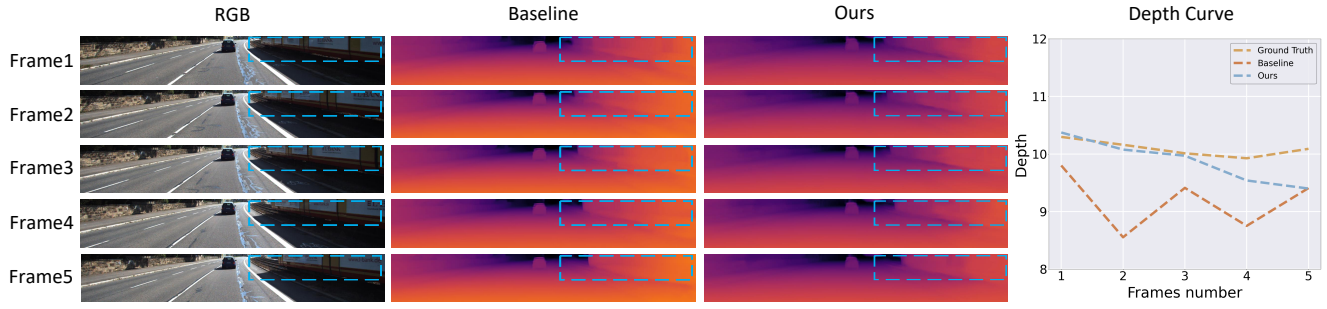

Figure 7: Qualitative depth results on the KITTI dataset [3]. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

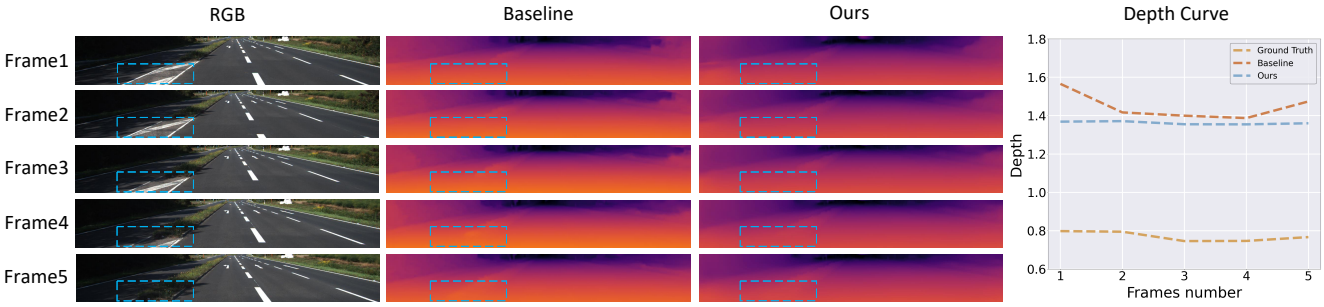

Figure 8: Qualitative depth results on the KITTI dataset [3]. We highlight regions with obvious difference in dashed rectangular. For better comparison, we draw depth curves on the last column. Each curve represents depth value for the center point of a certain box in the input frames.

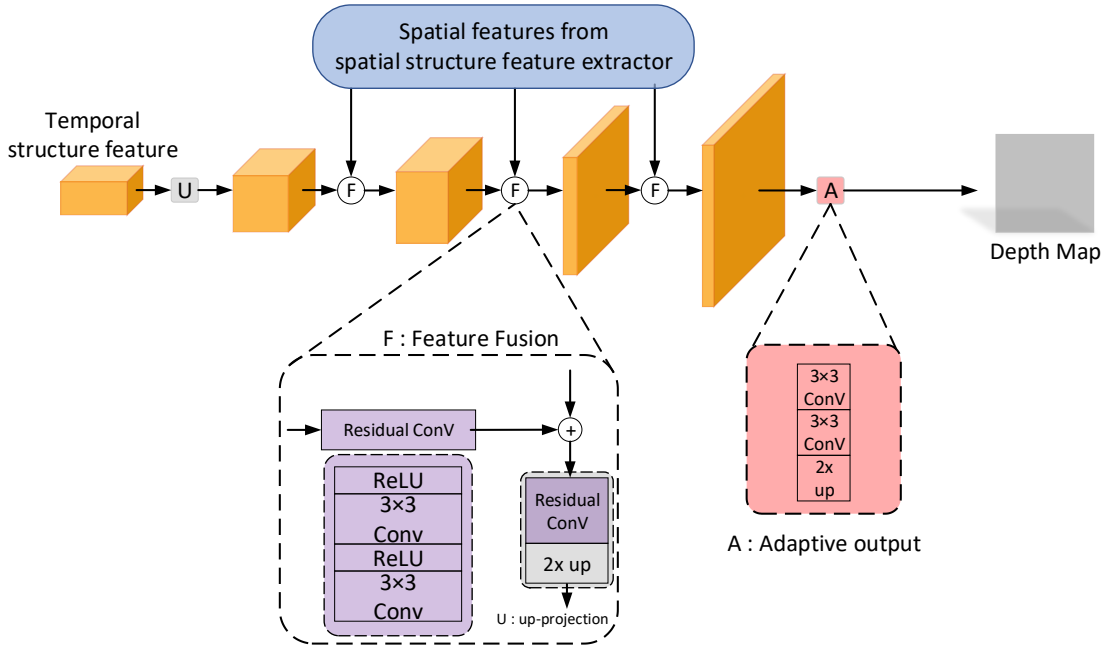

Figure 9: The network architecture of our depth predictor.