ACM Multimedia 2022
Lisbon, Portugal | 10-14 October

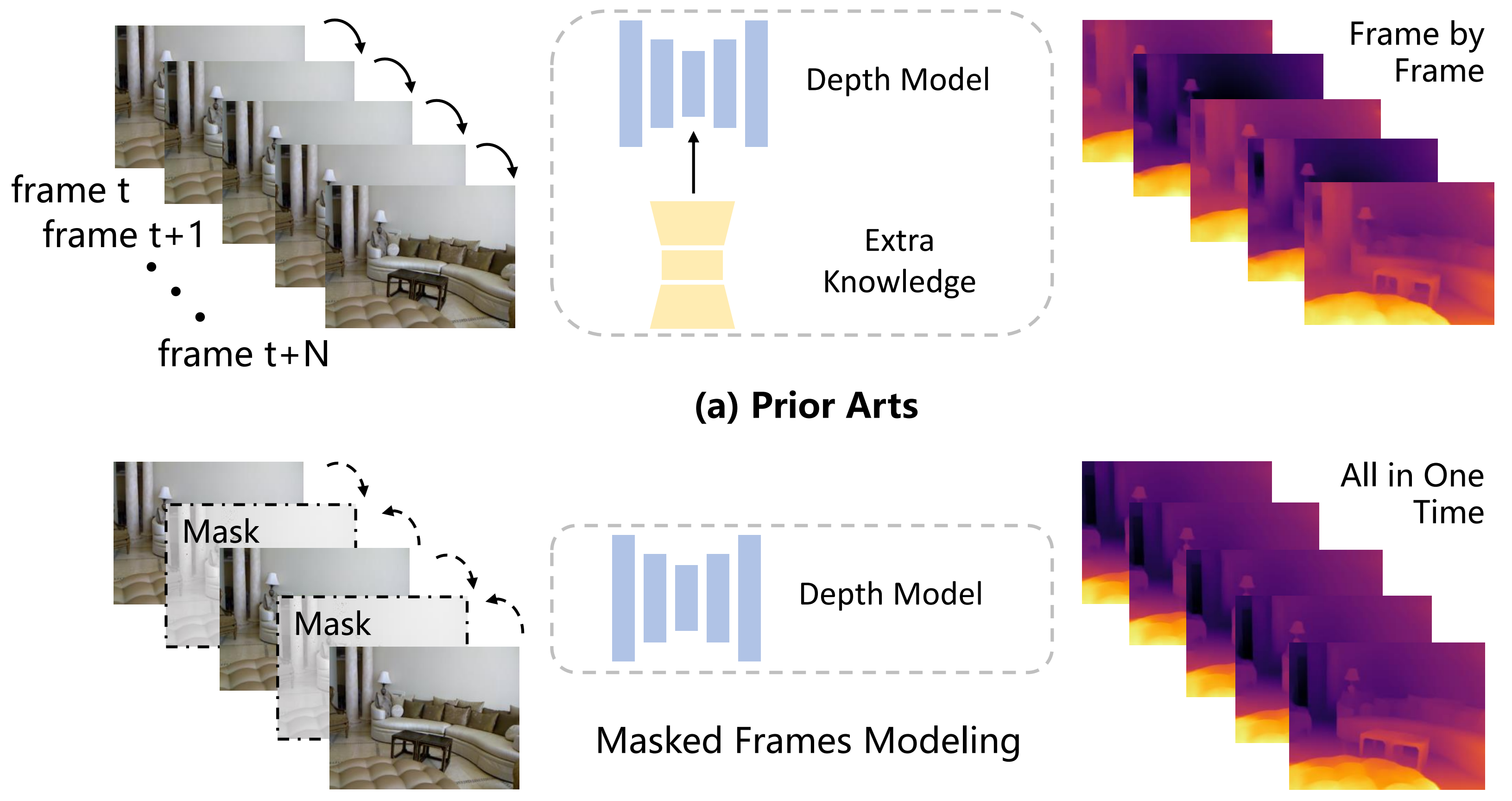
Less is More: Consistent Video Depth Estimation with Masked Frames Modeling

Yiran Wang[1], Zhiyu Pan[1], Xingyi Li[1], Zhiguo Cao[1], Ke Xian[1*], Jianming Zhang[2]

[1]School of AIA, Huazhong University of Science and Technology
[2]Adobe Research

Problem Statement

Goal: Our goal is to derive temporal consistency in video depth results. In some real-world applications, e.g., 2D-to-3D video conversion and video bokeh rendering, depth consistency plays a vital role.



(a) Prior Arts

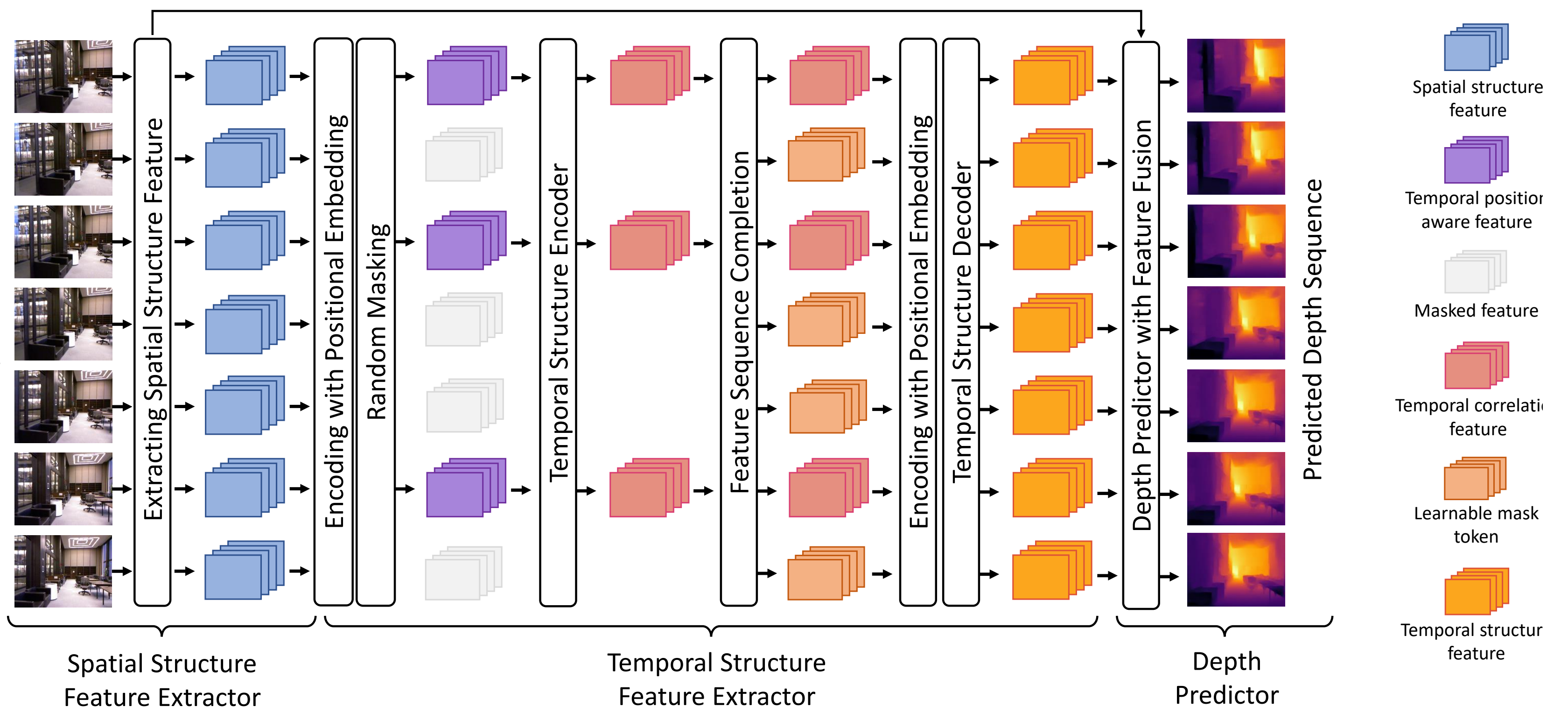
(b) Ours

Motivation: Previous methods model inter-frame correlations based on extra temporal clues. They could fail when those temporal clues are inaccurate. They are also inefficient and time-consuming.

Key Contributions:

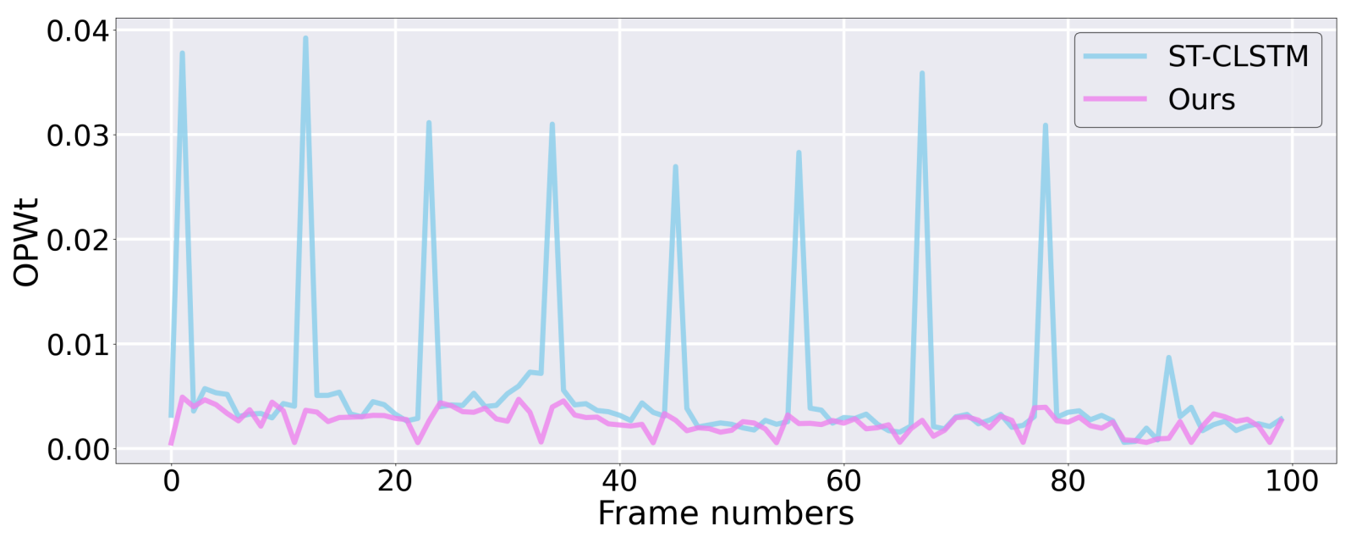
- We propose a masked video transformer for consistent video depth estimation without relying on optical flow, pose estimation, and GANs.
- To the best of our knowledge, we are the first to introduce ConvTransformer for video depth estimation, which can encode inter-frame correlations in parallel.

Framework


Extracting Spatial Structure Feature
Encoding with Positional Embedding
Random Masking
Temporal Structure Encoder
Feature Sequence Completion
Encoding with Positional Embedding
Temporal Structure Decoder
Depth Predictor with Feature Fusion
Predicted Depth Sequence
Spatial Structure Feature Extractor
Temporal Structure Feature Extractor
Depth Predictor
Spatial structure feature
Temporal positior aware feature
Masked feature
Temporal correlati feature
Learnable mask token
Temporal structur feature

Spatial Feature Extractor: generates the spatial feature map of each frame.
Temporal Structure Encoder: randomly masks a portion of input frames and encodes temporal correlation features.
Temporal Structure Decoder: reconstructs the features of masked frames according to the unmasked temporal correlation features.
Depth Predictor: recovers depth maps from the temporal structure features.

Motivation Review



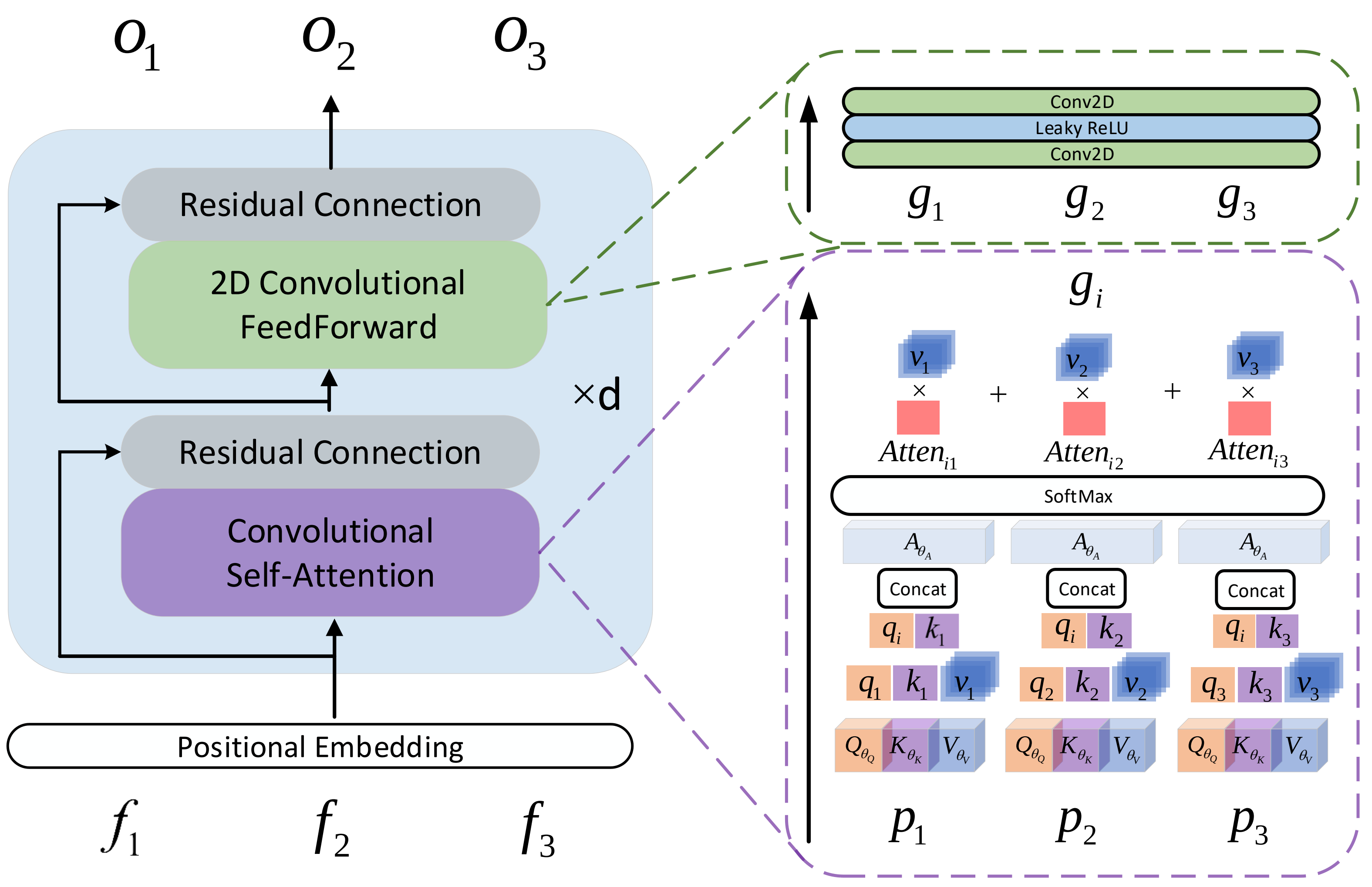
Method	Camera pose	Optical flow	Time(s)
Ours (ResNet18)			1.32
Ours (ResNext101)			3.36
DeepV2D (pose pre-computed)	✓		11.65
DeepV2D (pose estimation)	✓		55.62
CVD	✓	✓	376.32
Robust-CVD	✓	✓	252.17

Temporal Consistency: Our masked video transformer model has better characteristics of parallelism and globality. The larger temporal receptive field leads to better consistency whether inside or between input sequences without relying on additional optical flow or camera poses.

Inference Speed: We compare the inference speed of different methods. We can see that our FMNet, whether with ResNet18 or ResNext101 as the backbone, has significantly faster inference speeds than methods using pose estimation or optical flow such as DeepV2D, CVD, and Robust-CVD.

ConvTransformer

The original transformer is not appropriate for processing sequences of high-dimensional features due to the computational overhead of attention mechanism. We adopt the idea of ConvTransformer, which can directly process features sequences without partitioning patches and flatten operation.



Query, key, and value: generated by convolution sub-networks $\mathcal{Q}_{\theta_\mathcal{Q}}, \mathcal{K}_{\theta_\mathcal{K}}, \mathcal{V}_{\theta_\mathcal{V}}$

$$q_i = \mathcal{Q}_{\theta_\mathcal{Q}}(p_i)\,, k_i = \mathcal{K}_{\theta_\mathcal{K}}(p_i)\,, v_i = \mathcal{V}_{\theta_\mathcal{V}}(p_i)\,, i \in [0, N-1]\,.$$

Attention: approximated by convolutional sub-network $\mathcal{A}_{\theta_\mathcal{A}}$

$$Atten(i,j) = \mathcal{A}_{\theta_\mathcal{A}}(concat[q_i, k_j])\,, \quad i,j \in [0, N-1]\,,$$

Experiments

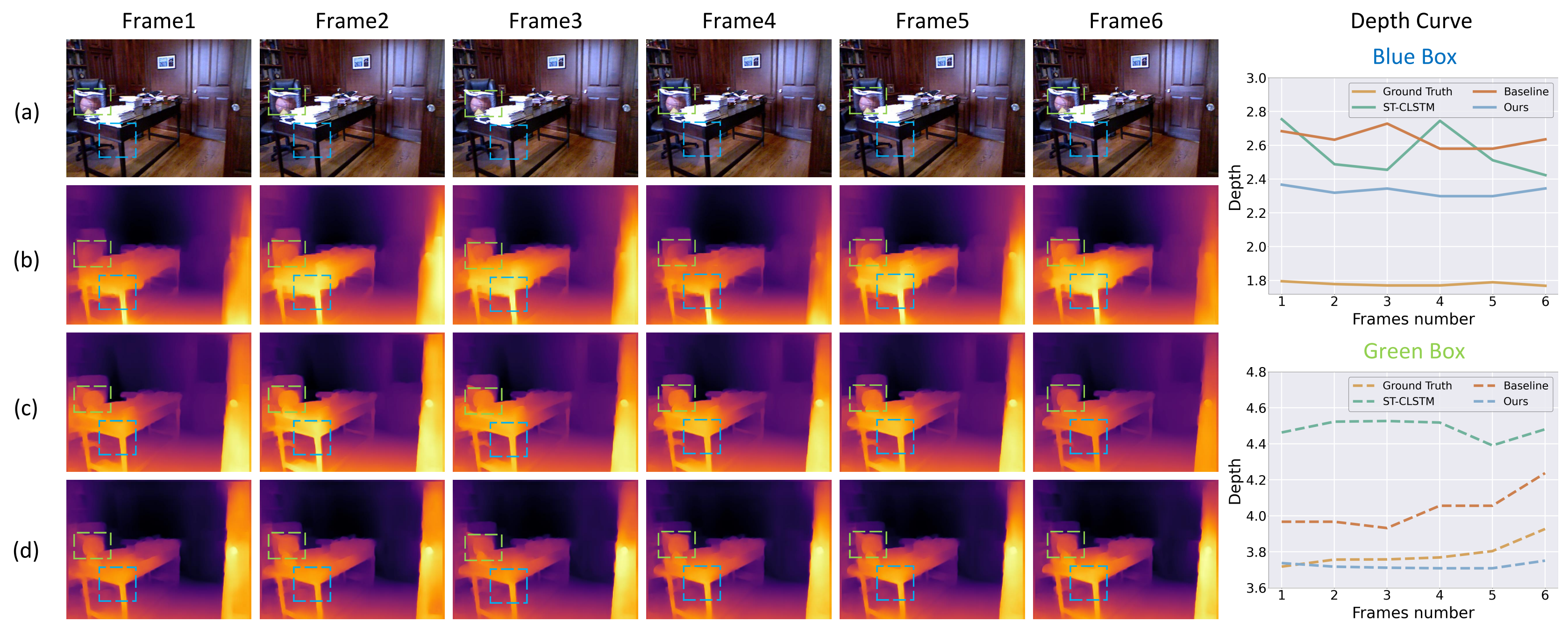
Quantitative results on the NYU Depth V2 dataset:

Method	Rel	RMSE	log 10	δ_1	δ_2	δ_3
Laina *et al.*	0.127	0.573	0.055	0.811	0.953	0.988
Pad-net	0.120	0.582	0.055	0.817	0.954	0.987
Cao *et al.*	0.141	0.540	0.060	0.819	0.965	0.992
DORN	**0.115**	0.509	**0.051**	0.828	0.965	0.992
ST-CLSTM	0.131	0.571	0.056	0.833	0.965	0.991
Cao *et al.*	0.131	0.574	0.056	**0.835**	0.965	0.990
Ours	0.134	**0.452**	0.056	0.832	**0.968**	**0.992**

Quantitative results on the KITTI dataset:

Method	Rel	RMSE	log 10	δ_1	δ_2	δ_3
Mahjourian *et al.*	0.159	5.912	–	0.784	0.923	0.970
Zhou *et al.*	0.143	5.370	–	0.824	0.937	0.974
ST-CLSTM	0.101	4.137	0.043	**0.890**	**0.970**	0.989
Patil *et al.*	0.111	4.650	–	0.883	0.961	0.982
CVD	0.130	4.876	–	0.878	0.946	0.970
Cao *et al.*	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**

Qualitative results on the NYU Depth V2 dataset:



Contact Information and Acknowledgement

Email: wangyiran@hust.edu.cn **Project:** https://github.com/RaymondWang987/FMNet **Acknowledgement:** This work was funded by Adobe Research