DeepVideoMVS: Multi-View Stereo on Video with Recurrent Spatio-Temporal Fusion

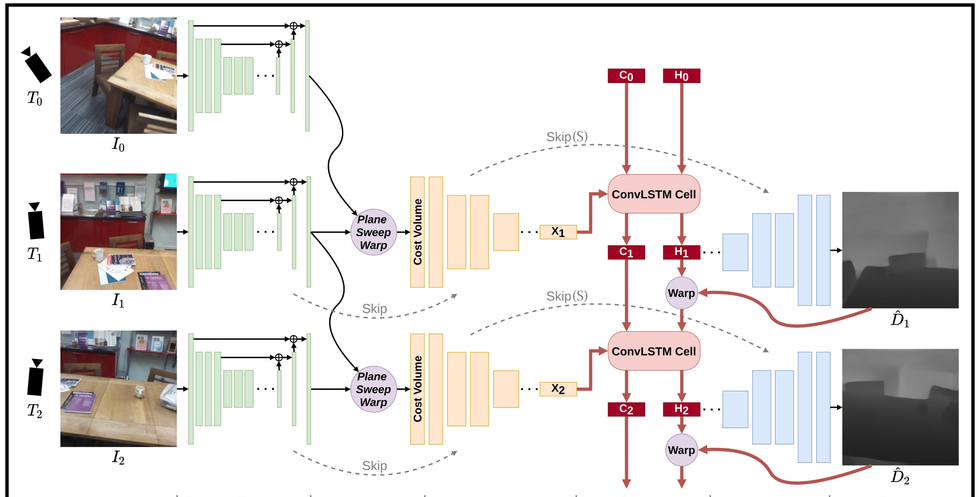
Arda Düzçeker[1] Silvano Galliani[2] Christoph Vogel[2] Pablo Speciale[2] Mihai Dusmanu[1] Marc Pollefeys[1,2]

[1]Department of Computer Science, ETH Zurich [2]Microsoft Mixed Reality & AI Zurich Lab

github.com/ardaduz/deep-video-mvs

Introduction

- Applications like navigation for autonomous vehicles and mixed reality require real-time access to the depth information.
- Visual data is often acquired as a video instead of sparse instances in time, and the depth is recovered for the selected keyframes.
- We propose a deep neural network that exploits the temporally structured input and harnesses the advantage of having limited viewpoint variation among the successive keyframes.
- The network combines:
  - A **lightweight** stereo depth estimation backbone.
  - A **memory cell** that acts as a fusion module and agglomerates the information obtained within the bottleneck.
  - A **hidden state propagation scheme** that addresses the viewpoint dependence of the encodings.
- DeepVideoMVS, **an online multi-view depth prediction approach**, predicts **consistent depth maps** and achieves **state-of-the-art depth prediction accuracy** while maintaining a **real-time** performance.



Hidden State Propagation Scheme

- Naïve fusion enables the information flow along time but forces the ConvLSTM cell to implicitly capture the pose-induced image motion between the input latent encoding and the hidden state.

Naïve Fusion

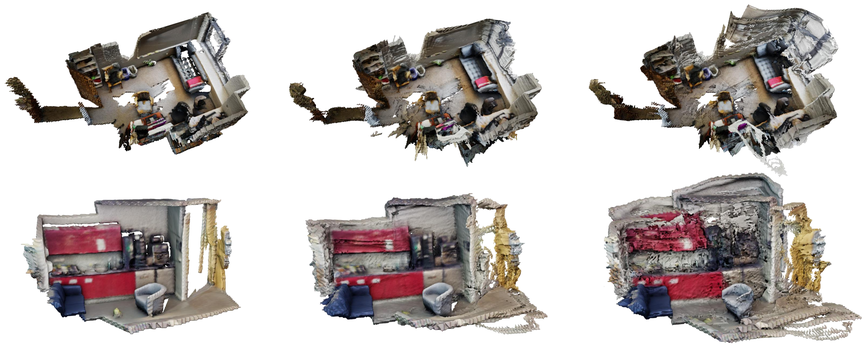
$$\mathbf{X}_t, \mathbf{S}_t = \text{encoding}(\mathbf{I}_t,\ \mathbf{I}_{t-1},\ \mathbf{T}_t,\ \mathbf{T}_{t-1},\ \mathbf{K})$$
$$\mathbf{H}_t, \mathbf{C}_t = \text{cell}(\mathbf{X}_t,\ \mathbf{H}_{t-1},\ \mathbf{C}_{t-1})$$
$$\hat{\mathbf{D}}_t = \text{decoding}(\mathbf{H}_t,\ \mathbf{S}_t).$$

Proposed Fusion

$$\widetilde{\mathbf{D}}_t = \text{projection}(\hat{\mathbf{D}}_{t-1},\ \mathbf{T}_{t-1},\ \mathbf{T}_t,\ \mathbf{K})$$
$$\widetilde{\mathbf{H}}_{t-1} = \text{warping}(\mathbf{H}_{t-1},\ \widetilde{\mathbf{D}}_t)$$
$$\mathbf{X}_t, \mathbf{S}_t = \text{encoding}(\mathbf{I}_t,\ \mathbf{I}_{t-1},\ \mathbf{T}_t,\ \mathbf{T}_{t-1},\ \mathbf{K})$$
$$\mathbf{H}_t, \mathbf{C}_t = \text{cell}(\mathbf{X}_t,\ \widetilde{\mathbf{H}}_{t-1},\ \mathbf{C}_{t-1})$$
$$\hat{\mathbf{D}}_t = \text{decoding}(\mathbf{H}_t,\ \mathbf{S}_t).$$

- In proposed fusion, the hidden state is warped to the next time-step's viewpoint with bilinear grid sampling.
- Alleviates the need to capture the flow of the visual representations.
- To set the training in motion and stabilize the behaviour, groundtruth depth maps are used for warping.

Backbone vs. Proposed Fusion

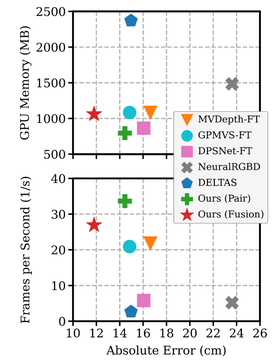


Groundtruth **With Proposed Fusion** **Backbone Only**

- Our fusion approach improves the quality and the consistency of the depth predictions with minimal overhead. This results in significantly less noisy reconstructions that are also geometrically more accurate.

Experiments and Results

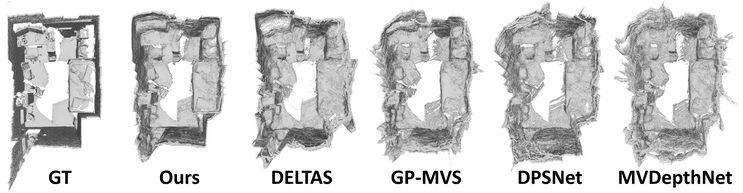
- Supervised training on ScanNet's training split and evaluation on five different indoor datasets.
- Propagating the hidden state as is, brings 7.4% improvement over the backbone in absolute depth error, while the proposed propagation leads to 19.3%.
- Our fusion approach outperforms the best competitor, finetuned GP-MVS, by 20.3% in absolute inverse depth error.
- Our fusion approach has the highest frame rate and consumes less memory than the most.



Speed and memory consumption in relation to the depth prediction performance on the ScanNet test split.

	MVDep	MVDep (FT)	DPSNet	DPSNet (FT)	DELTAS	Ours (Pair)	NRGBD	GPMVS	GPMVS (FT)	Ours (Fusion)
SCANNET										
abs	0.1953	0.1671	0.2185	0.1607	0.1492	0.1454	0.2361	0.2027	0.1491	**0.1187**
abs-inv	0.0621	0.0540	0.0710	0.0530	0.0506	0.0468	0.0745	0.0669	0.0490	**0.0381**
7-SCENES										
abs	0.2029	0.2012	0.2486	0.1966	0.1911	0.1860	0.2143	0.1962	0.1737	**0.1448**
abs-inv	0.0732	0.0708	0.0847	0.0729	0.0717	0.0653	0.0756	0.0756	0.0641	**0.0537**
RGB-D V2										
abs	0.1387	0.1310	0.1520	0.1320	0.1581	0.1433	**0.1227**	0.1576	0.1275	0.1256
abs-inv	0.0617	0.0637	0.0763	0.0623	0.0811	0.0667	0.0700	0.0761	**0.0559**	0.0566
TUM RGB-D										
abs	0.2902*	0.3260	0.2958*	0.3045	0.3525	0.3535	0.3185	**0.2443***	0.2938	0.2878
abs-inv	0.0643*	0.0715	0.0758*	0.0722	0.0753	0.0736	0.0681	0.0604*	0.0643	**0.0551**
ICL-NUIM										
abs	**0.1392**	0.1574	0.1695	0.1491	0.1953	0.1771	0.1730	0.1667	0.1558	0.1496
abs-inv	0.0305	0.0328	0.0367	0.0359	0.0458	0.0402	0.0365	0.0356	0.0323	**0.0297**

Performance on: **i.** ScanNet test set, **ii.** 13 sequences from 7-Scenes , **iii.** 8 sequences from RGB-D Scenes V2, **iv.** 13 sequences from TUM RGB-D and **v.** 4 sequences from Augmented ICL-NUIM. *Bold* is the best score, *overline* indicates the second best score. The vertical line separates video agnostic (*left*) from video aware (*right*) methods. * the method is already trained on most of the test frames.



GT **Ours** **DELTAS** **GP-MVS** **DPSNet** **MVDepthNet**

Conclusion

- We present a framework –that can extend many existing MVS methods– to improve the depth outputs when processing video streams.
- We show that our fusion approach combined with our lightweight backbone delivers accurate depth maps while being real-time capable.