DynamicStereo: Consistent Dynamic Depth from Stereo Videos Supplementary Material

Nikita Karaev[1,2] Ignacio Rocco[1] Benjamin Graham[1] Natalia Neverova[1]
Andrea Vedaldi[1] Christian Rupprecht[2]

[1] Meta AI [2] Visual Geometry Group, University of Oxford

Attention	Sintel Clean		Dynamic Replica	
	Bad3px	TEPE	Bad1px	TEPE
None	6.47	0.779	7.31	0.119
Space + Stereo	6.14	0.814	7.37	**0.116**
Space + Stereo + Time	**6.02**	**0.753**	**5.50**	0.120

Table 1. **SST-Block Attention.** We compare no attention (none), spatial (space) and time attention.

Attention	Sintel Clean		Dynamic Replica	
	Bad3px	TEPE	Bad1px	TEPE
None	6.42	0.940	7.41	0.117
Space	**5.99**	0.864	7.26	**0.114**
Space + Time	6.02	**0.753**	**5.50**	0.120

Table 2. **Update Block Attention.** A combination of space and time attention helps to propagate information.

1. Additional Ablations

We ablate the proposed SST block and the choice of attention in the update block. We train the model on SceneFlow [3] for 50k iterations with the same hyper-parameters as in the main paper.

SST Block Attention We evaluate the choice of attention types of the SST-Block in Tab. 3. We find that including attention layers generally improves disparity estimation both in terms of accuracy and temporal consistency. Attention across space, stereo pairs and time achieves the best results. Interestingly, time attention also improves accuracy, potentially through the use of multiple viewpoints over time improving the precise location of correspondences.

Update Block Attention In Tab. 2 we compare different choices of attention inside the Update Block. The model with a combination of space and time attention performs well on both datasets. Similarly, improvements are gained in both stereo and temporal metrics.

2. Implementation details

Here we provide additional implementation details.

Training For all the DR & SF dataset generalization experiments, we sample the same number of *frames* from both DR and SF. For temporal consistency experiments, we sample the same number of *sequences* from DR and SF.

We found that learnable positional encoding for time can generalize better during inference on longer sequences. We thus use learnable encoding for time and sin / cos Fourier features for space.

Augmentations During training, we set image saturation to a value sampled uniformly between 0 and 1.4. We stretch right frames to simulate imperfect rectification: it is stretched by a factor sampled uniformly from $[2^{-0.2}, 2^{0.4}]$. Following [5], we simulate occlusions by randomly erasing rectangular regions from each frame with probability 0.5.

Inference For better temporal consistency during inference, we split the input video into 20-frame chunks with an overlap of 10 frames. We then apply the model to each chunk and discard the first and the last 5 frames of each prediction to compose the final sequence of disparity estimations.

Space-Stereo-Time attention We add time and position encoding to left and right input feature tensors and reshape them to $(B*T, \frac{H}{16}*\frac{W}{16}, d)$. Then we apply linear self-attention [4] across space to both tensors and cross attention across space between left and right tensors. Finally, we reshape left and right tensors to $(B*\frac{H}{16}*\frac{W}{16}, T, d)$ and apply standard attention across time.

3D CNN-based GRU For efficiency, each 3D GRU module is composed of three separable height-width-time GRUs with kernel sizes $(1 \times 1 \times 5)$, $(5 \times 1 \times 1)$, and $(1 \times 5 \times 5)$.

Upsampling To pass the output of each update block g to a higher resolution update block, we use a combination of convex upsampling from RAFT [5] and standard bi-linear upsampling.

Method	sec./frame
RAFT Stereo [2]	0.83
CODD [1]	1.04
DynamicStereo (Ours)	1.20

Table 3. **Runtime analysis.** We run each method on a video of resolution 1280x720 on GPU and report the average number of seconds it takes the method to process one frame.

3. Limitations

While our method is more temporally consistent than previous work, it still is not fully stable over time. This partially comes form the fact that the method is evaluated in a sliding window fashion resulting in low frequency oscillations at the scale of the window size (1-2 sec). Extending the window size is currently not possible due to memory limitations.

As any stereo matching method, exceedingly large untextured scene parts such as walls and other surfaces are difficult to predict accurately. Learning from DynamicReplica helps to learn priors to mitigate this issue but does not solve it completely.

As dense groundtruth information is near impossible to collect, evaluation and training relies on synthetic datasets such as DynamicReplica. Generalization to the real world can only be assessed qualitatively and might not fully reflect the performance on artificial scenes.

References

[1] Zhaoshuo Li, Wei Ye, Dilin Wang, Francis X Creighton, Russell H Taylor, Ganesh Venkatesh, and Mathias Unberath. Temporally consistent online depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3018–3027, 2023. 2

[2] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *International Conference on 3D Vision (3DV)*, 2021. 2

[3] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. 1

[4] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1

[5] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow (extended abstract). In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4839–4843. ijcai.org, 2021. 1