Structure and Motion from Casual Videos

Supplementary Materials

1 An Overview Video

We provide a narrated video for our paper in the supplementary material, which highlights the key contributions of our paper. It also contains 3D visualizations of our results.

2 Additional Ablations on Sintel

Table 1. Results for additional experiments on Sintel.

	Static (Rel. L1)	Dynamic (Rel. L1)	All (Rel. L1)	ATE	RRE
MiDaS	0.305	1.468	0.697	—	—
RCVD	0.427	1.505	0.847	0.164	1.151
Ours w. Binary Mask	0.281	2.268	0.765	0.135	0.713
Ours w. Zhou et al.[5] Mask	0.392	3.088	1.155	0.269	5.905
Ours, full	**0.227**	**1.267**	**0.484**	**0.089**	**0.410**

We provide additional movement map ablations in this section, reporting quantitative depth results separately for the static and dynamic part of the Sintel dataset.

First, we ablate our method using a binary movement mask(Tab.1, *Ours w. Binary Mask*), using the straight-through estimator for differentiability [1].

We then compare with the movement estimation strategy by Zhou et al. [5], which treats the movement map as weights for the L1 loss and adds a sparsity prior to avoid trivial solutions. We empirically find the sparsity prior quite sensitive and fail to generate good results. Here we report the result using the default value (0.5) from Zhou et al. [5] (Tab.1, *Ours w. Zhou et al. [5]* Mask).

3 DAVIS: Full Results

We provide full qualitative results on the DAVIS dataset for both our method and Robust-CVD [2]. The result is in video format and shows depth, motion segmentation and residual flow results for both methods. Due to the size limit of supplementary materials, 3D visualizations of the results are not included. We are happy to host them anonymously upon reviewer request, if the guidelines permit.

As shown in Sec. 4.3 in the main paper, our method largely performs well on DAVIS. However, we observed several cases where the results are suboptimal. These include situations where the depth of moving objects is inaccurate (8 videos), videos featuring changing intrinsics over the course of the video (3 videos), videos with fisheye lenses (4 videos), videos where initial depth is inaccurate (1 video), and situations where the movement map switches between foreground and background (3 videos). Overall we consider 71 of 90 videos to be successes, with 79 partial successes (camera poses are accurate, but depth maps are not), and 11 failures. As mentioned in the main paper, there are several avenues to improve these results, including integrating recent work in moving object depth prediction [4].

Video results are provided separately for successful tracks, tracks with inaccurate depth of moving objects, tracks with changing intrinsics, tracks shot with fisheye lenses, tracks with bad depth initialization and tracks where movement map switches.

4 Network Details

We here provide the details of the networks used in our work. For the depth network, we only optimize the `refine` layers of the MiDaS V2 [3] decoder, which consist of 8 residual convolution layers. For the movement map decoder, we use the second-to-last feature map from the MiDaS V2 encoder, and decode it with a light-weight network. Assuming the input image is of size $H \times W$, the network specifics are:

ID	Operation	Output Shape
Input Image	–	$3 \times H \times W$
Encoder Feature	–	$256 \times H/16 \times W/16$
C1	3×3 Conv	$128 \times H/16 \times W/16$
N1	InstanceNorm 2d	$128 \times H/16 \times W/16$
A1	ReLU	$128 \times H/16 \times W/16$
U1	Bilinear Upsample	$128 \times H/8 \times W/8$
C2	3×3 Conv	$128 \times H/8 \times W/8$
N2	InstanceNorm 2d	$128 \times H/8 \times W/8$
A2	ReLU	$128 \times H/8 \times W/8$
U2	Bilinear Upsample	$128 \times H/4 \times W/4$
C3	3×3 Conv	$64 \times H/4 \times W/4$
N3	InstanceNorm 2d	$64 \times H/4 \times W/4$
A3	ReLU	$64 \times H/4 \times W/4$
U3	Bilinear Upsample	$64 \times H/2 \times W/2$
C4	3×3 Conv	$32 \times H/2 \times W/2$
N4	InstanceNorm 2d	$32 \times H/2 \times W/2$
A4	ReLU	$32 \times H/2 \times W/2$
U4	Bilinear Upsample	$32 \times H \times W$
C5	3×3 Conv	$1 \times H \times W$
A5	ELU	$1 \times H \times W$

Table 2. Details of the movement map decoder. The decoder takes in the feature map from the MiDaS feature encoder, and outputs a single channel movement map with the same spatial resolution as the input image.

References

1. et al., B.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
2. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
3. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
4. Zhang, Z., Cole, F., Tucker, R., Freeman, W.T., Dekel, T.: Consistent depth of moving objects in video. ACM Transactions on Graphics (TOG) **40**(4), 1–12 (2021)
5. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)